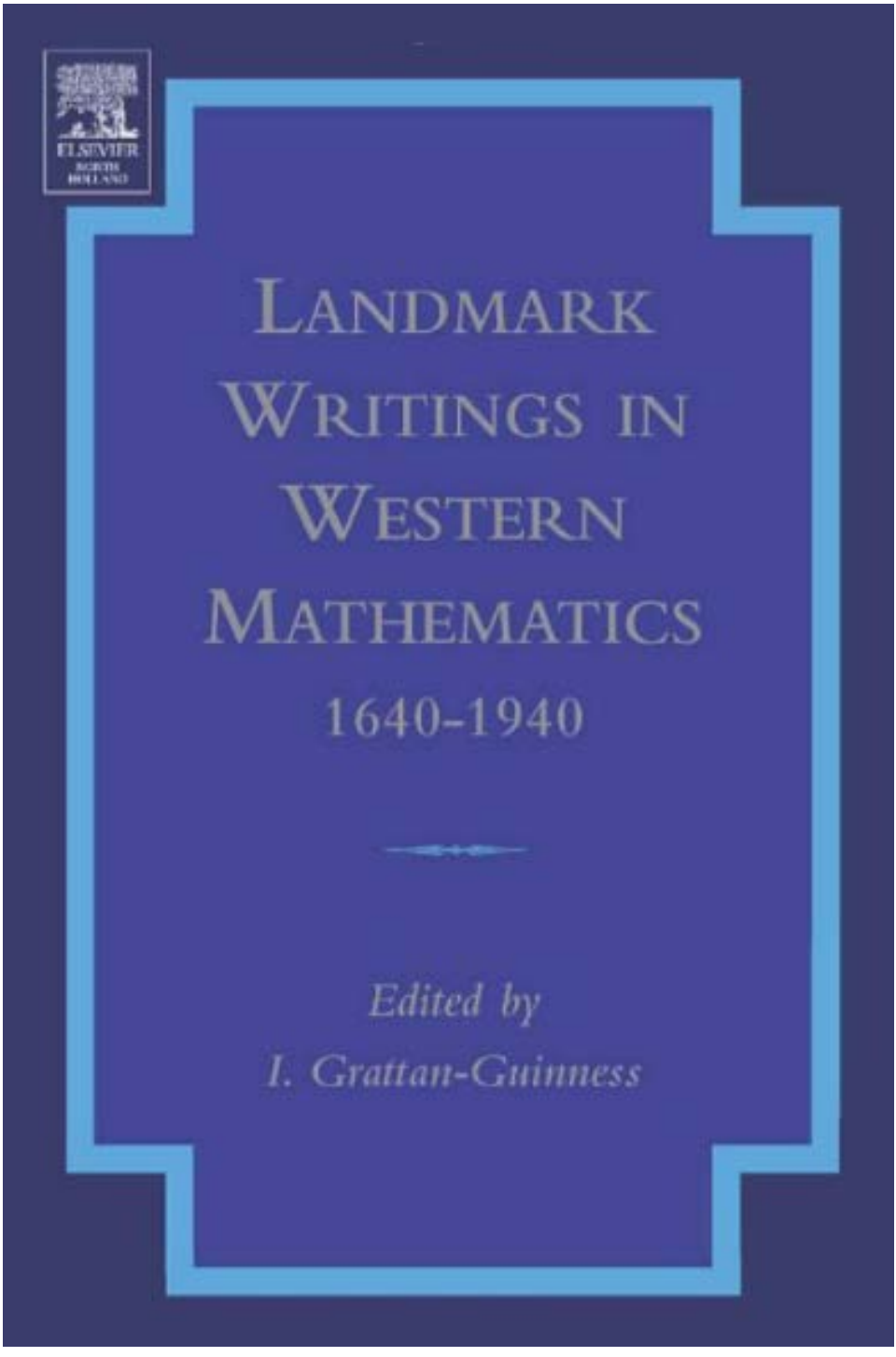

ELSEVIER
NORTH HOLLAND
LANDMARK WRITINGS IN WESTERN MATHEMATICS
1640-1940
Edited by
I. Grattan-Guinness

Landmark Writings in Western Mathematics 1640–1940

Edited by

I. Grattan-Guinness

Editorial Board

Roger Cooke
Leo Corry
Pierre Crépel
Niccolo Guicciardini

2005

Amsterdam • Boston • Heidelberg • London • New York • Oxford •
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo

TABLE OF CONTENTS

In many cases a short title or a description of the writing is given. The date is that of first publication; the first edition is involved unless otherwise indicated.

CHAPTER 0

INTRODUCTION

I. Grattan-Guinness

1 WAVES IN THE SEA

For a very long time mathematical research has been circulated as a stream of books and papers, manuscripts and letters; in ancient times scrolls and tablets prevailed, and in recent ones emails and electronic files have joined in. Most writings duly took or take their modest or perhaps overlooked place in the flow; but some have made a major impact on the branches and aspects of mathematics to which they refer, and maybe also to other branches and even disciplines not originally within their purview. This book is devoted to a substantial number of the principal writings of this kind that were published during the period 1640–1940. The order of articles is that of the appearance of the writings involved: they are cited throughout the book by article number, in the manner '§21'. The table of contents indicates the range of topics to be covered; this introduction explains the scope and limitations of the choice of writings, and the manner of their treatment.

Usually the text discussed is a book; but sometimes it is one or more papers in a journal, such as N.H. Abel on the quintic equation in 1826 (§29). Thus the more general word 'writing' is used to describe the chosen text. This book is composed of 'articles', which are divided into 'sections' (whereas some of the original writings fall into 'Sections'). The articles contain cross-references to other articles; '§21.3' refers to section 3 of article 21.

Most articles deal with one writing each; but in a few cases more than one are taken together when they handle closely related topics and were published within a short time; for example, G.W. Leibniz launching his version of the calculus in three papers between 1684 and 1693 (§4), or two books by different pairs of authors in the mid 1930s (§76). A multivolume work is considered in total even if the spread of time is as great as that needed by P.S. Laplace to assemble his mathematical astronomy (1796–1827: §18). Normally the first edition of a book catches the attention; but the second (1799–1802) edition of Etienne Montucla's history of mathematics is taken in §21, for it covered a wider range and enjoyed much more impact than did the first edition of 40 years earlier.

2 ORGANISATION OF THE ARTICLES

Each article begins bibliographically, with the publication history of the writing(s) as far as we have been able to track it down: first publication, later reprints and/or editions where

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

applicable, and (photo)reprints and translations. (In the case of modern reprints that may have been reprinted several times themselves, such as with publishers such as Dover and Chelsea, we have given only the date of the *initial* reprinting.) We exclude short extracts or parts of a writing as reproduced in general anthologies. When known, the location of the manuscript of the original writing is recorded. Finally, cross-references to related articles are listed.

In the article proper the career of the author is briefly reviewed, with especial attention given to the place of the writing in his career. Its prehistory and content are surveyed, and for writings of some length a table of contents is supplied, divided up into suitable units such as parts or chapters. Page numbers are usually given, either of the first page number of each unit or the number of pages that it occupies; unless indicated otherwise, they pertain to the original printing. Then principal features of the writing are described and discussed, and on occasion omissions of topics that one might expect to have seen handled. In a few cases a portrait of an author is included, and the title page of a writing if it includes a nice design, say, or carries an interesting motto. Some original diagrams are included.

Then comes the impact of the writing—often the most difficult part of the history to assess. When impact was made fairly quickly, authors have concentrated upon the first 30–40 years or so, including where appropriate upon the later work of its author(s); striking cases of negative influence are noted. But for several writings the reception was quite tardy—the ripples from Hermann Grassmann's *Ausdehnungslehre* (1844) took nearly 40 years to propagate before being gaining a quantity of admirers, for instance (§32)—and such facts are noted, and where possible explanations are suggested. When different parts of the writing received rather different impacts, each one is discussed *in situ*. On literary style, some authors write in the past ('Newton showed') while others adopt the historic present ('Newton shows').

Some writings have appeared in various printings, editions and translations: thus several page numbers are involved. Where practical, passages in the writing are cited by article or chapter numbers. Otherwise the original printing is cited unless it has become very rare; for example, George Green's book of 1828 on potential theory (§30), where the reprint in the edition of his works is cited.

Each article ends with a bibliography of relevant items, mostly historical ones but some primary ones also. Several items, sometimes all of them, are cited in the text, in the style '[Smith, 1976]'. Reprints of these items are often indicated, especially in editions of works. Other items, especially of primary literature, are usually mentioned by name in the article, and with sufficient precision for the reader to be able to track them down.

The book ends with a list of affiliations of the authors and their articles, and an index.

3 SOME PRINCIPAL LIMITATIONS

3.1 Period

The chosen period begin around 1640, when mathematics (and science in general) was beginning to show the first signs of professional employment and diffusion of information as we know it; for example, somewhat more publication than before, the founding the

Royal Society of London and the *Académie des Sciences* in Paris in the 1660s, and the launch of scientific journals such as the *Acta Eruditorum* which was Leibniz's venue in 1684 and later. Some comments need to be made about the immediate pre-history.

In research up to and during the early 17th century, geometry was Euclidean, but the range of curves and surfaces had expanded far beyond the repertoire deployed in Euclid's *Elements*. Topics included methods of determining tangents to curves or surfaces and areas enclosed by them, in methods now often called 'pre-calculus'. Partly in these connections some functions and series were developed; also various numerical methods, especially logarithms. Algebra was much concerned with properties of polynomial equations. Mechanics was a major concern, usually with different theories obtaining in the terrestrial and celestial branches (the latter including technology of machines and artefacts). Trigonometry, planar and spherical, was part of the mathematical wardrobe, especially for cartography with navigation and astronomy, though its heyday for research was largely over. Much less developed topics include probability theory and mathematical statistics, and number theory. Professional support was modest; universities were best developed in Italy. Another important type of employer was the leader of a country or state. Few mathematicians made a living from their work; for several their research was a hobby.

Major figures from the immediately preceding generation include Johannes Kepler in Germany, Galileo Galilei (died 1642) in Italy, Simon Stevin in the Netherlands, and John Napier and Thomas Harriot in Britain. They and other contemporaries and immediate forerunners come at the end of a *different* European story, which begins with the transmission of ancient and Arabic sources into Europe in the 11th and 12th centuries and the translation into Latin of most of them, and then the reliance upon manuscripts being supplemented and later overtaken by the introduction of printing from the late 15th century onwards. That story is substantially different from ours, and needs a separate book; to encompass both would require more space than is available here. The same remark could be made about the history of writings elsewhere in the world, such as in the Far East.

The publication terminus of 1940 is chosen not only because of the Second World War but also the massive size of candidate later writings. Various survey books or encyclopaedias on branches of recent mathematics can be consulted: for example, [Pier, 1994, 2000], largely for pure mathematics.

3.2 Choice of the writings

It would have been easy but rather tedious and narrow-minded to dominate the list of writings with a procession of undoubtedly major treatises on mathematical analysis, algebras, mechanics and mathematical physics. A principal purpose of this book is to exhibit the *range* and *variety of theories* within mathematics as it has developed over the period covered. Thus writings have been selected from both pure and applied mathematics, including probability and statistics, and their selection was guided by their global place as well as by their intrinsic merits: for example, that a writing was not only important but also is the only representative of its rather unusual area (such as Stanley Jevons on mathematical economics in §41).

It was decided to have *articles* on the chosen writings, and not create a much more numerous list of dictionary-like summaries. This policy made selection even more severe;

77 articles cover 89 writings from across the mathematical board of the period. Cut down from an original list of more than twice the length, several of the final choices were difficult to make, and omissions do not entail criticism. The selected ensemble offers, we hope, a reasonable characterisation of the full ensemble. If Your Favourite Writing is missing, dear reader, then we mortals have offended.

Within the full sequence of articles occur some sub-sequences of articles on writings in and around the same branch of mathematics or topic. Table 1 indicates the main such sub-sequences.

The book is not offered as a *general history* of mathematics, even for its period. For various important developments have taken place without any one writing being significant enough to have gained an article here. Among many examples, Newton's 'fluxional' version of the calculus gradually became known in manuscript form from the late 1660s onwards and then in print from 1704, but no one (or even two) versions are sufficiently significant to enter our roll; however, Colin MacLaurin's *Treatise on fluxions* (1742) is the subject of §10. Again, Karl Weierstrass's lectures at the University of Berlin cast a huge influence upon his students and their own later endeavours for nearly 30 years from the late 1850s; but none was published at that time, and the line of influence from any one of them is too tenuous to be described, or to be highlighted over those of the other lecture courses. Among branches of mathematics, numerical methods are not well represented, as they have not generated major writings in our sense; however, several methods are mentioned in some articles.

Another criterion for selection was that the impact of the writing had to be reasonably (inter)national. This required that it be written in a widely read language or soon translated into one, or at least that much of its contents became known well beyond its geographical origin. Among writings that did not meet this criterion, some Russian works have been casualties, in particular several excellent Soviet achievements.

In some cases the impact of a writing was so late that the achievement involved was acknowledged as anticipation and maybe as a general inspiration for later work but not as an active source for its prosecution. For example, Leonhard Euler's solution in 1736 of the Königsberg bridge problem was a remarkable pioneering effort in graph theory and combinatorics; but it does not seem to have led to the development, much later, of both subjects [Biggs and others, 1976], and so is not given an article here.

Another excluded source is a short statement. For example, Pierre de Fermat's conjecture in number theory, which became known rather optimistically as his 'last theorem', was posthumously published in 1679; but his few lines involved have not been taken as a writing as such.

Also excluded are all manuscripts (including letters) that were published only much later or not (yet) at all; for while the achievements in them may have been remarkable, no broad impact was made. However, manuscripts pertinent to the history of a chosen writing (such as its own manuscript, as mentioned above) are noted in some articles.

3.3 Mathematical level

In the early 1820s A.-L. Cauchy launched his version of real-variable mathematical analysis, a landmark process indeed (§25). However, the writings involved are very unusual for

Table 1. Groupings of writing by principal branches of mathematics. Often a shortened title or indicative description is used.

Geometries	**Algebras**
1649 Descartes, *Geometria* (§1)	1649 Descartes, *Geometria* (§1)
1744 Euler on curves (§12)	1826 Abel on the quintic equation (§29)
1748 Euler, *Introductio* to analysis (§13)	1844 Grassmann, *Ausdehnungslehre* (§32)
1795 Monge, *Géométrie descriptive* (§17)	1853 Hamilton, *Lectures on quaternions* (§35)
1822 Poncelet on projective geometry (§27)	1854 Boole, *Laws of thought* (§36)
1844 Grassmann, *Ausdehnungslehre* (§32)	1863 Dirichlet, *Vorlesungen über Zahlentheorie* (§37)
1847 von Staudt, *Geometrie der Lage* (§33)	1872 Klein, Erlangen programme (§42)
1867 Riemann on geometries (§39)	1895–1896 Weber, *Lehrbuch der Algebra* (§53)
1872 Klein, Erlangen programme (§42)	1897 Hilbert on algebraic number fields (§54)
1899 Hilbert, *Grundlagen der Geometrie* (§55)	1930–1931 van der Waerden, *Moderne Algebra* (§70)
1905–34 Enriques and Chisini on algebraic geometry (§62)	
Calculus	**Number theory**
1684–93 Leibniz, first papers on the calculus (§4)	1801 Gauss, *Disquisitiones arithmeticae* (§22)
1734 Berkeley, *The analyst* (§8)	1863 Dirichlet *Vorlesungen über Zahlentheorie* (§37)
1742 MacLaurin, *Treatise on fluxions* (§10)	1897 Hilbert on algebraic number fields (§54)
1744 Euler on curves (§12)	1919–1923 Dickson, *Number theory* (§65)
1755 Euler, *Differentialis* (§14)	
1797 Lagrange, *Fonctions analytiques* (§19)	
1797–1800 Lacroix, *Traité du calcul* (§20)	
Functions, series, differential equations	**Real and complex analysis**
1656 Wallis, *Arithmetica infinitorum* (§2)	1823 Cauchy, *Résumé* of the calculus (§25)
1748 Euler, *Introductio* to analysis (§13)	1825, 1827 Cauchy, two main writings on complex analysis (§28)
1797 Lagrange, *Fonctions analytiques* (§19)	1851 Riemann on complex analysis (§34)
1797–1800 Lacroix, *Traité du calcul* (§20)	1867 Riemann on trigonometric series (§38)
1799–1827 Laplace, *Mécanique céleste* (§18)	1904 Lebesgue, *Intégration* (§59)
1821 Cauchy, *Cours d'analyse* (§25)	1932 Bochner on Fourier integrals (§74)
1822 Fourier on heat diffusion (§26)	
1829 Jacobi, *Functionum ellipticarum* (§31)	**Set theory, foundations**
	1872 Dedekind, *Stetigkeit und Irrationalzahlen* (§43)

Table 1. (*continued*)

1905–06 Baire on discontinuous functions and Lebesgue on trigonometric series (§59)	1883 Cantor, *Grundlagen* of set theory (§46)
1932 Bochner on Fourier integrals (§74)	1888 Dedekind, *Was sind Zahlen?* (§47)
	1889 Peano on axioms for arithmetic (§47)
General mechanics	1910–1913 Whitehead and Russell, *Principia mathematica* (§61)
1687 Newton, *Principia* (§5)	1931 Gödel's incompletability theorem (§71)
1743 d'Alembert, *Dynamique* (§11)	1934, 1939 Hilbert and Bernays, *Grundlagen der Mathematik* (§77)
1788 Lagrange, *Méchanique analitique* (§16)	
1867 Thomson and Tait, *Treatise on natural philosophy* (§40)	**History, general**
1894 Hertz, *Prinzipien der Mechanik* (§52)	1799–1802 Montucla, *Histoire des mathématiques* (§21)
Astronomy	1892 Rouse Ball, *Mathematical recreations* (§50)
1687 Newton, *Principia* (§5)	1901 Hilbert, paper on mathematical problems (§57)
1788 Lagrange, *Méchanique analitique* (§16)	
1799–1827 Laplace, *Mécanique céleste* (§18)	**Dynamics**
1809 Gauss, *Theoria motus* (§23)	1673 Huygens, *Horologium* (§3)
1890 Poincaré on the three-body problem (§48)	1738 Daniel Bernoulli, *Hydrodynamica* (§9)
	1890 Poincaré on the three-body problem (§48)
Probability and statistics	1893 Lyapunov, *Stability theory* (§51)
1713 James Bernoulli, *Ars conjectandi* (§6)	1927 Birkhoff, *Dynamical systems* (§68)
1718 De Moivre, *Doctrine of chances* (§7)	
1764 Bayes on probability theory (§15)	**Mathematical physics**
1809 Gauss, *Theoria motus* (§23)	1822 Fourier on heat diffusion (§26)
1812–1814 Laplace, *Probabilités* (§24)	1828 Green, *Electricity and magnetism* (§30)
1854 Boole, *Laws of thought* (§36)	1844 Grassmann, *Ausdehnungslehre* (§32)
1900 Pearson on the chi-squared test (§56)	1873 Maxwell, *Electricity and magnetism* (§44)
1925 Fisher, *Statistical methods* (§67)	1877–1878 Rayleigh, *Theory of sound* (§45)
1931 Shewhart, *Economic quality control* (§72)	1892 Heaviside, *Electrical theory* (§49)
1933 Kolmogorov on the foundations of probability theory (§75)	1904 Thomson, *Baltimore lectures* (§58)
	1909 Lorentz on electrons (§60)

Table 1. (*continued*)

Topology	1916 Einstein on general relativity theory (§63)
1889 Poincaré on the three-body problem (§48)	1930 Dirac, *Quantum mechanics* (§69)
1923–1926 Urysohn and Brouwer on dimensions (§66)	1932 von Neumann, *Quantenmechanik* (§69)
1934 Seifert and Threlfall, *Topologie* (§76)	
1935 Alexandroff and Hopf, *Topologie* (§76)	**Social and life sciences**
	1871 Jevons, *Theory of political economy* (§41)
	1917 Wentworth Thompson, *On growth and form* (§64)
	1931 Volterra on mathematical biology (§73)

this volume in being two textbooks; they were chosen because much of their content was new—and, as is revealed in the article, the students hated them! The writings for this book have been chosen for their *research content*: the novelties that they contained and/or 'the state of the art' which they comprehensively summarised. A comparable review of textbooks at various stages of education requires a companion volume, for a massive literature of its own is involved—in terms of print-runs, often much larger than those of the writings discussed here. While historians of mathematical education will find some material of interest here, the main audience is historically sympathetic mathematicians, members of kindred disciplines, and historians of mathematics.

3.4 Journals

Since the majority of research mathematics appears in journals, then their own inaugurations constitute landmarks. However, they have not been treated here, since they embody a different kind of history. It is not well covered: Erwin Neuenschwander provides a valuable short survey in [Grattan-Guinness, 1994, art. 11.12].

3.5 Landmarks as epitaphs

The writings discussed here either launched new phases of work, or consolidated the known state of theory on a topic, of both. But theories and traditions sometimes die, or at least die down: for example for mathematics, the last fluxional textbooks in the 1810s and 1820s, the fading away of quaternions (for a long time anyway) in the early 20th century, or the final calculations of massive invariants of high order. This volume bears upon declines only when a chosen writing treats theories that were soon to be noticeably eclipsed. An arresting example is Lord Kelvin's 'Baltimore lectures' of 1884 on aspects of classical mathematical physics (§58), which were fully published in 1904, just before the emergence of Albert Einstein.

ACKNOWLEDGEMENTS

A history of this 'Great Books' type has not been attempted for mathematics before, or maybe for any science; I am indebted to Arjen Sevenster of Elsevier for suggesting that the time was ripe for a try, and to him and to Mrs. Andy Deelen for bearing the brunt of the publisher's consequences. I thank the authors not only for preparing their articles but also for helping me and each other with all sorts of details; and my editorial board for advice on the choice of writings and of authors, and for reading articles. I am indebted to Dr. Ben Garling for agreeing to let me complete the editing of the article on Cauchy on complex-variable analysis by Frank Smithies (§28), who sent me the final version on the day before he died in November 2002. Those articles written in French or German were kindly translated into English by Dr. D.L. Johnson of Nottingham University, and checked with the authors.

BIBLIOGRAPHY

In addition to the items cited in the text above, this list includes the main general biographical and bibliographical sources for the history of mathematics.

Books

Biggs, N., Lloyd, E.K. and Wilson, R.J. 1976. *Graph theory 1736–1936,* Oxford: Clarendon Press. [Slightly revised re-issue 1998.]

Dauben, J.W. (ed.) 1984. *The history of mathematics from antiquity to the present. A selective bibliography,* New York: Garland. [2nd ed., ed. A.C. Lewis: Providence, RI: American Mathematical Society, 2000, CD-ROM.]

Gowers, W.T. and Barrow-Green, J. (eds.) To appear. *The Princeton companion to mathematics*, Princeton: Princeton University Press.

Gottwald, S., Ilgauds, H.-J., and Schlote, K.-H. (eds.) 1990. *Lexikon bedeutender Mathematiker,* Leipzig: Bibliographisches Institut. [Wide range of capsule biographies; rather weak on engineering mathematicians.]

Gillispie, C.C. (ed.) 1970–1980. *Dictionary of scientific biography,* 16 vols., New York: Scribners. [Vols. 17–18 (1990) ed. F.L. Holmes.]

Grattan-Guinness, I. (ed.) 1994. *Companion encyclopedia of the history and philosophy of the mathematical sciences*, 2 vols., London: Routledge. [Longer general bibliography in article 13.1. Repr. Baltimore: Johns Hopkins University Press, 2003.]

Heyde, C.C. and Seneta, E. (eds.) 2001. *Statisticians of the centuries*, New York: Springer.

May, K.O. 1973. *Bibliography and research manual in the history of mathematics*, Toronto: University of Toronto Press.

Pier, J.-P. (ed.) 1994. *Development of mathematics 1900–1950,* Basel: Birkhäuser.

Pier, J.-P. (ed.) 2000. *Development of mathematics 1951–2000,* Basel: Birkhäuser.

Serials

British Society for the History of Mathematics, *Newsletter,* 1986–2003, *Bulletin*, 2004–. [Contains a section of abstracts in each issue.]

Historia mathematica, San Diego: Academic Press, 1974–. [Contains a large section of abstracts in each issue.]

Jahrbuch über die Fortschritte der Mathematik, 68 vols., Berlin: de Gruyter, 1867–1942. [The abstracting journal of its period; massive coverage.]
Mathematical reviews, 1940–, Providence, RI: American Mathematical Society. [Reasonable coverage of the historical literature since around 1973.]
Zentralblatt für Mathematik und ihre Grenzgebiete, Berlin: Springer, 1931–.

Internet

Many writings in all fields are available for downloading from some websites; for example, 'gallica', maintained by the *Bibliothèque Nationale* in Paris.
A valuable site with links to many other sites is run by the British Society for the History of Mathematics at www.dcs.warwick.ac.uk/bshm/.
A site devoted to general and biographical information in the history of mathematics is maintained by St. Andrew's University at the url: http://turnbull.mcs.st-and.ac.uk/history/.
Another site with more specialized focus on mathematics in culture is at Simon Fraser University: http://www.math.sfu.ca/histmath.

This page intentionally left blank

CHAPTER 1

RENÉ DESCARTES, *GÉOMÉTRIE*, LATIN EDITION (1649), FRENCH EDITION (1637)

M. Serfati

Inspired by a specific and novel view of the world, Descartes produced his *Géométrie*, a work as exceptional in its contents (analytic geometry) as in its form (symbolic notation), which slowly but surely upset the ancient conceptions of his contemporaries. In the other direction, this treatise is the first in history to be directly accessible to modern-day mathematicians. A cornerstone of our 'modern' mathematical era, the *Géométrie* thus paved the way for Newton and Leibniz.

First publication. *La Géométrie* (hereafter, 'G37'), Leiden: Jan Maire, 1637 [at the end of the *Discours de la Méthode* ('*DM*')]. 118 pages.

Latin editions. 1) *Geometria* (trans. Frans van Schooten), Leiden: Maire, 1649, x + 118 pages. [With *Notae Breves* de F. de Beaune, a commentary by van Schooten and a *Additamentum*.] 2) Three further editions, in 1659–1661 ('G59'), Amsterdam: 1683 ('G83') and Frankfurt/Main, 1695. 3) Derivatives of G59, with commentaries of Van Schooten, de Beaune, J. Hudde (*De reductione equationum*), J. de Witt (*Elementa curvarum*) and H. Van Heuraet. [G59 and G83 are available on http://gallica.bnf.fr/.]

Principal French editions. 1) Paris: C. Angot, 1664. 2) Ed. V. Cousin in *Oeuvres de Descartes*, Paris: Levraut, 1824–1826, vol. 5. 3) In *Oeuvres de Descartes* (ed. C. Adam and P. Tannery), vol. 6, Paris: L. Cerf, 1896, 367–485. [Common today, frequently reprinted and re-edited; the source of citations here. Hereafter, '[ATz]', where 'z' is the Roman number of the volume: see further in the bibliography.]

English translation. *The geometry of René Descartes* (ed. D.E. Smith and M.L. Latham), London and Chicago: Open Court, 1925. [Repr. New York: Dover, 1954. Includes a facsimile of G37.]

German translation. *Die Geometrie* (ed. L. Schlesinger), 1st ed., Leipzig: Mayer and Müller, 1894. [4th ed. 1923; repr. Darmstadt: Wissenschaftliche Buchgesellschaft, 1969.]

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

Italian translation. La geometria, in *Opere scientifiche*, vol. 2 (ed. E. Lojacono), Turin: UTET, 1983.

Related articles: Wallis (§2), Leibniz (§4), Newton (§5), Euler *Introductio* (§13), Lacroix (§20).

1 YOUTH, FROM LA FLÈCHE TO THE *REGULAE*

The main biographical source, the *Vie de M. Descartes* by Father Baillet [1691], is sometimes unreliable; so we supplement it with the critical biography [Rodis-Lewis, 1995]. Born in 1596 at La Haye en Touraine (now 'Descartes') on the borders of Poitou into the minor aristocracy, the young René Descartes (then 'Du Perron') studied at the Jesuit college of La Flèche, near le Mans. Following the educational reforms of Christopher Clavius, school mathematics was taught only as a subsidiary subject. After leaving La Flèche, Descartes took a degree in law at Poitiers and then travelled to Holland to enlist as a soldier under the orders of Maurice de Nassau (Prince Maurice of Orange).

It was at Breda in November 1618 that he met Isaac Beeckman (1588–1637) in front of a poster displaying a mathematical problem. The Dutch scientist, eight years older than Descartes, was impassioned with 'physico-mathematics', a new concept with a name coined by him. Having formed a friendship with Beeckman, Descartes was pleased to discover in him a perspective on science different from contemporary esoteric theories such as that of Raimond Llull. Descartes gave him his first scientific work, the *Compendium musicae* [ATx, 79–141]. On 26 March 1619, he enthusiastically sent him an important letter [ATx, 165–160], in which he concluded that 'there is almost nothing left to discover in geometry' ('*adeo ut pene nihil in Geometria supersit inveniendium*') and that he possessed the elements of 'a completely new science' ('*scientia penitus nova*') and had discovered 'a light to dispel the deepest darkness' ('*luminis* [...] *cujus auxilio densissimas quasque tenebras discuti posse existimo*').

Descartes then went to Germany, to the Duchy of Neuburg. On the night of 10 November 1619, shut in a heated room (an 'oven'), he had three dreams (see [ATx, *Olympica*]), which he always declared to be decisive in determining his scientific vocation and his Method [Rodis-Lewis, 1995, 60–71]. The Cartesian commentary is traditionally divided into two periods, before and after the 'oven'.

Having left his 'oven', Descartes decided to make a radical change in his way of life and to 'get rid of his prejudices' garnered from scholastic philosophy [Baillet, 1691]. He left for 'nine years of exercise in the Method', a wandering life of which the timetable is not known, except for a visit to Italy in 1623–1625; during it he made no attempt at publication. Following his return to Paris in 1625, he was a member of the Mersenne group until the death of the latter. Marin Mersenne (1588–1648), a religious Minim, was a recognized scientist. The 'reverend father', competent and clever, was for a long time the 'secretary of European science', a scientific intermediary between scholars with Descartes at the head (see the letters in [ATi–ATv]). This was also the time of Descartes's 'Rules for the direction of the mind' [Descartes, 2002], a posthumous text not intended for publication. It was based on the mathematical method, and central to the elaboration of the Cartesian philosophy of science: the *Géométrie* depended directly on it, even more than on the *DM*. Rule IV, for

example, develops various privileges of the method of discovery: what matters is not what is found without method, 'by chance', but rather why and how it is found. Descartes also explains the superiority of analysis as he conceived it (one starts with the effect or result, assuming it to have been achieved) over synthesis.

The objective was not, however, the advancement of mathematics, but a description of the physical world using the *Mathesis universalis*, a method of analysis based on mathematical procedures. In the Cartesian view of an objective world, where space was identified with his 'solidified' geometric structure, geometry and mechanics were only to play the role of sciences. And the reduction of the physical world to the geometric was thus essential, but the geometry of the measurable. Centred around 'order and measure', the *Regulae* thus proposed a modern view of the 'real numbers'. Rule XVI contains the first exponent in history ('$2a^3$'), a Cartesian invention, with the consequence that a^2 could represent a line, and not necessarily an area. This position opposed ancient conceptions of the world from the Greek dualism (numbers versus magnitudes) to the Cossic system of 'species' (squares, cubes, and so on) and also Franciscus Vieta.

This was a decisive step. In the Cartesian doctrine, according to the *Regula*, algebra existed only as a tool useful for stating and solving geometric problems. Not being a formalist, Descartes had no interest in symbolic problems, and it was somehow in spite of himself that he became a founder of modern mathematical symbolism [Serfati, 1994]. For him the role of algebra lay chiefly in the register of 'mechanical imagination' [Rodis-Lewis, 1995], which he loved: he believed that the automatic nature of calculation enabled the mathematician to free himself from relying on his memory.

This period came to an end in 1628 through a meeting in Paris with Bérulle, a papal legate. Descartes decided to emigrate to Holland, where he spent nearly all the rest of his life, frequently changing his address.

2 DESCARTES IN HOLLAND AND STOCKHOLM

In the autumn of 1628, Descartes met Beeckman briefly at Dordrecht. He also began, around 1630, the drafting of a treatise on physics, *Le Monde* or *Traité de la lumière*, in which he defended his position on the movement of the earth; but he discontinued the work for prudence following the condemnation of Galileo in 1633. In 1631, Jacobus Golius (1596–1667), professor of mathematics and oriental languages at Leyden, submitted to him the problem of Pappus (discussed in section 6 below).

In 1637, Descartes put out at Leiden the *DM* and the *Essais*, of which *Géométrie* was the only mathematical work he published. By then he was a philosopher of European reputation: scholars and theologians argued over his philosophy. The year 1640 saw the death of his daughter Francine, 'the greatest sadness I have ever known'. The Cartesian philosophy was established with the publication, in Paris and then in Amsterdam (1641 and 1642), of the 'Meditations on the first philosophy' (*Meditationes de prima philosophia*) along with the 'Objections and replies' (*Objectiones cum responsionibus authoris*).

The second period in Holland (1642–1649) was marked for Descartes by various difficulties caused by the accusations of atheism made against his philosophy by the Dutch universities and theologians (compare the quarrel of Utrecht in [Rodis-Lewis, 1995, 227–243]. It was punctuated by important publications, such as the 'Principles of philosophy'

(*Principia philosophiae*), which was actually a treatise on Cartesian physics, published in Amsterdam in 1644; and G49, published by Frans van Schooten (1615?–1660) in 1649.

In 1649, Descartes received an invitation from Queen Christine of Sweden to visit Stockholm and teach her his philosophy. With little enthusiasm, he set off in September 1649, but died a few months later, on the morning of 11 February 1650, in Stockholm at the age of 54 [Rodis-Lewis, 1995, 261–297]. The manuscripts he took with him make up the 'Stockholm inventory' [ATx, 5–12].

3 THE *GÉOMÉTRIE*

The *Géométrie* is the last of the *Essais* (that is, texts of application) in the *DM*. The four texts were published in sequence (418 + 31 pages) without the name of the author on 8 June 1637 by the printing-house of Jan Maire in Leyden, under the title 'Discourse on the Method of correct reasoning and the search for truth in the sciences, then the *Dioptrics*, the *Meteors* and the *Geometry*, which are tests of this method' ('*Discours de la Méthode pour bien conduire sa raison, et chercher la vérité dans les sciences, plus la* Dioptrique, *les* Météores, *et la* Géométrie, *qui sont des Essais de cette Méthode*'). *DM* is a celebrated philosophical text, in which the *Géométrie* is the only mathematical application. It contains, besides an autobiography, a description of the four principles of the 'Méthode' anchored in the practice of mathematics and the preference for 'thinking clearly and distinctly'.

Descartes finished the drafting of the *Géométrie* at the very last moment, making the final discoveries as the *Météores* was being printed. The *Essais* were written in French in such a way that non-specialists and 'even ladies' could understand them. The *Géométrie* was ignored in several editions of the *DM*. In contrast, it appeared on its own in the Latin edition G49 (our 'Landmark' writing), which was aimed at the European public (Figure 1); it also contained a commentary by van Schooten and a translation of the notes sent to Descartes by Florimond de Beaune (1601–1652), an admirer of the *Géométrie*, soon after its appearance. Descartes was aware of the originality of the project: 'Finally, in the *Géométrie* I try to give a general method for solving all the problems that have never been solved'. He was aware of the difficulty of the work, and asked the reader to 'follow all the calculations, which may seem difficult at first, with pen in hand', whereupon he will get used to them 'after a few days'. He also advised passing 'from the first book to the third before reading the second' [ATi, 457–458].

4 THE 'CONSTRUCTION' OF THE EXPRESSIONS

The contents of the *Géométrie/Geometria* are summarised in Table 1. It is divided into three Books, but the layout does not follow a coherent scheme. This paradox in a 'methodical' setting, as is noted in the commentary [Bos, 1991], was undoubtedly a secret strategy adopted by Descartes for several reasons [Serfati, 1993]. A main one was to disguise the true depth of his methods in such a way that his enemies, such as Gilles de Roberval (1602–1675), whom he hated, could not imagine after the event that they know them before Descartes. The work is thus difficult to read. Three 'threads of Ariadne' can


GEOMETRIA,
à
RENATO DES CARTES
Anno 1637 Gallicè edita; nunc autem
CUM NOTIS
FLORIMONDI DE BEAVNE,
In Curiâ Blæſenſi Conſiliarii Regii,
In linguam Latinam verſa, & commentariis illuſtrata,
Operâ atque ſtudio
FRANCISCI à SCHOOTEN,
Leydenſis, in Academiâ Lugduno-Batavâ, Matheſeos Profeſſoris, Belgicè docentis.
ACADLVGD
LVGDVNI BATAVORVM,
Ex Officinâ IOANNIS MAIRE,
cIↃ IↃC XLIX.

Figure 1.

be distinguished: the acceptability of curves, the problem of Pappus and the construction of roots ‘using curves’. Descartes describes every geometrical result in two phases, by ‘construction–demonstration’, that is, statement of the result and details of the proof, the latter being sometimes in another place or absent altogether. We shall adopt a double pagination, separated by the symbol ≡: the first number refers to [ATvi] and the second to G49.

Table 1. Contents of Descartes's book. The page numbers refer to the editions indicated. The abbreviations are explained in the text.

1637 (ATvi)	**1649 (G49)**	**Contents**
Book I *Planar problems. Ruler and compass*.		
369	1	I-A Geometrical constructions of algebraic expressions.
372	4	I-B Strategy of putting into equation of a geometrical problem. 'Construction' of planar problems.
377	8	I-C Statement of the problem of Pappus. Expression of the 'distance' from a point to a straight line. Equation for four lines.
Book II *The two criteria for the acceptability of curves.*		
388	19	II-A Acceptability according to the CM criterion. Cartesian compasses.
392	23	II-B Acceptability according to the algebraic criterion. Classification into species.
393	24	II-C Return to CM. Ruler and slide. The CP, image of a parabola.
396	28	II-D Return to the problem of Pappus in three or four lines. Reconnaissance of loci. Point-by-point constructions. Planar and solid loci. The CP, Pappus curve in five lines.
412	45	II-E Tangents and normals.
424	57	II-F Ovals.
440	74	II-G Notions for the case of three dimensions.
Book III *Algebraic equations. Constructions* (*by curves*) *of roots*.		
442	75	III-A Constructions of roots by auxiliary curves. 'Simplicity' of the methods.
444	77	III-B Roots of algebraic equations.
454	86	III-C Transformation and reduction of equations.
464	95	III-D Solid problems. 'Construction' of the third and fourth degrees by circle and parabola. Mean proportionals doubles. Trisection.
476	107	III-E 'More than solid' problems. [End $485 \equiv 118$.]

Book I begins with several definitions of notation, indispensable to the reader of the time, for whom the contents was new. This symbolic preamble was so important to Descartes that he had written up part of it ['Calcul de M. Descartes', ATx] before relinguishing the idea of publishing it separately. He listed the usual five operations, including the extraction of roots, and (like Vieta) introduced a segment of unit length to ensure the underlying physical homogeneity of written expressions: thus, $aabb - b$ must be understood as $aabb - b111$ (pp. $371 \equiv 3$). Letters always stand for lengths of segments, or positive numbers, just as with Pierre de Fermat (1607–1665). Following Vieta, Descartes used let-

ters a, b or x, z to represent known or unknown lengths in the same way (replacing the distinction between vowels and consonants by that between the beginning and end of the alphabet). Along with the Cartesian exponents (z^3 or b^4), of which this was the first public use, the symbols for the basic operations comprise such expressions as $a+b$, ab and $\sqrt{aa+bb}$. Like Vieta, Descartes described how to 'construct' them geometrically with ruler and compass. Thus, multiplication is done using a construction with parallel lines, the theorem of Thales and a unit segment, according to which $x/a = b/1$. Thus, 'ab' is represented, according the *Regulae*, by a line and not by the area of a rectangle: a decisive point in the Cartesian numerization of the world.

Descartes also introduced the specific symbol $\propto$ for equality. The notation is close to that used today. He explained a specific strategy (see below) for expressing a geometrical problem as an equation. He completely solved 'plane' problems, that is, the equation $z^2 = az + b$, both geometrically by ruler and compass and algebraically.

5 COMPASSES, RULER-AND-SLIDE, CRITERION FOR 'CONTINUOUS MOTIONS'

The first thread of Ariadne is to delineate the frontier between the curves acceptable in geometry, which Descartes called 'geometric', and the rest, which he called 'mechanical'; the modern terms 'algebraic' and 'transcendental" are due to G.W. Leibniz (1646–1716). The question had been open from antiquity (from Plato to Pappus), constructibility by ruler-and-compass usually defining the boundary. Being critical of Pappus's categorization of plane, solid and linear curves, Descartes gives in the *Géométrie* two criteria of acceptability.

Firstly, for Pappus 'plane' curves were those constructible by ruler and compass, 'solid' curves were the conic sections, and 'linear' curves were the rest, such as the conchoids, the spiral, the quadratrix and the cissoid. The linear curves were also called 'mechanical' by the ancient Greeks because instruments were needed to construct them. Observing that ruler and compasses are also instruments, Descartes extended this classification by introducing his set-square compasses. This is a mechanical system of sliding set squares that 'push' each other, the whole motion being regulated by the aperture of the compasses (Figure 2). A specifically Cartesian invention, these theoretical machines (he never constructed them) served him from his youth in the solution of cubic equations, the insertion of proportional means, and the construction of curves [Serfati, 1993; Serfati, 2003].

The compass-curves appear in Book II: in modern terms, the nth curve has the polar equation $\rho = a/(\cos\theta)^{2n}$, represented by dotted lines in Figure 2. They constitute a model conforming to the first criterion for 'continuous motions' (hereafter, 'CM'). In Descartes's words (pp. 389–390 ≡ 20–21):

> One must not exclude [from geometry] the composite lines any more than the simple ones, provided one can imagine them being described by a continuous motion, or by several such motions in sequence where the later ones are entirely governed by their predecessors.

On the other hand, Descartes excludes the quadratrix and spiral from the 'geometrical' curves since they are the result of separate motions, circular and rectilinear, without a con-

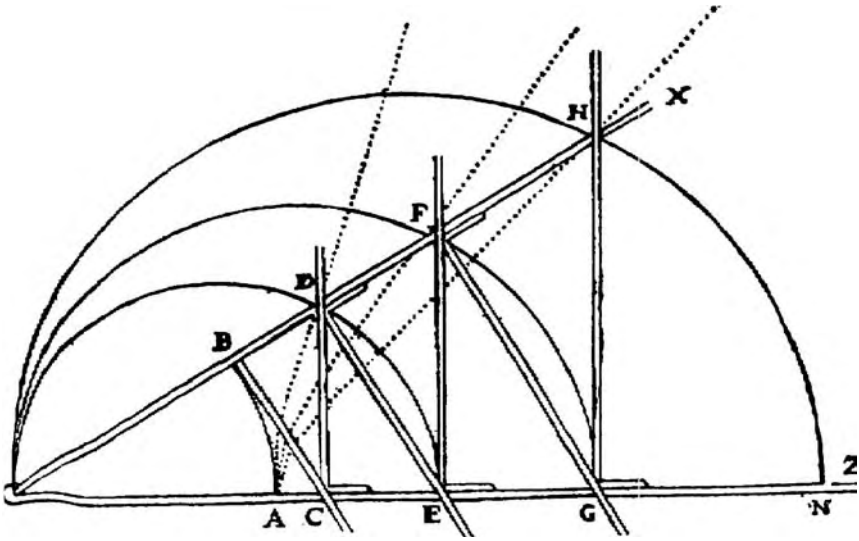


Figure 2. Set-square compasses.

nection 'that one could measure exactly'. Being dependent on a single integer parameter n, however, compasses could only describe a restricted class of curves. To extend it, Descartes introduced, always within the confines of CM, another articulated theoretical instrument, the ruler-and-slide (Figure 3).

This instrument enables one to associate to any curve 'embedded' in a moveable plane an 'image curve' in a fixed plane (compare Descartes's transformation in [Serfati, 2003]). To a line there thus corresponds a hyperbola and to a circle, a conchoid. To a parabola there corresponds a new curve 'one degree more complicated than the conic sections', crucial in the rhetoric of the *Géométrie* and called a 'Cartesian parabola' (hereafter, 'CP') by the historians Gino Loria and Henk Bos. It has the equation

$$axy = y^3 - 2ay^2 - a^2y + 2a^3 \left[= (y+a)(y-a)(y-2a)\right], \tag{1}$$

and belongs to a family of cubics that Isaac Newton (1642–1727) called 'tridents' in the *Enumeratio*: (1) is no. 66 in his classification. Such kinematically generated curves were not considered by Fermat, but they arise (in a different way) in the work of Roberval.

6 THE PROBLEM OF PAPPUS AND AN ALGEBRAIC CRITERION

In 1631, Golius sent Descartes a geometrical problem, that 'of Pappus on three or four lines', and this forms our second thread of Ariadne (for 'line' here, read 'straight line'). It had originally been posed and solved shortly before the time of Euclid in a work called *Five books concerning solid loci* by Aristaeus, and was then studied by Apollonius and later by Pappus. The solution was lost in the 17th century (see Paul Tannery's note on the problem of Pappus in [ATvi, 721–725]).

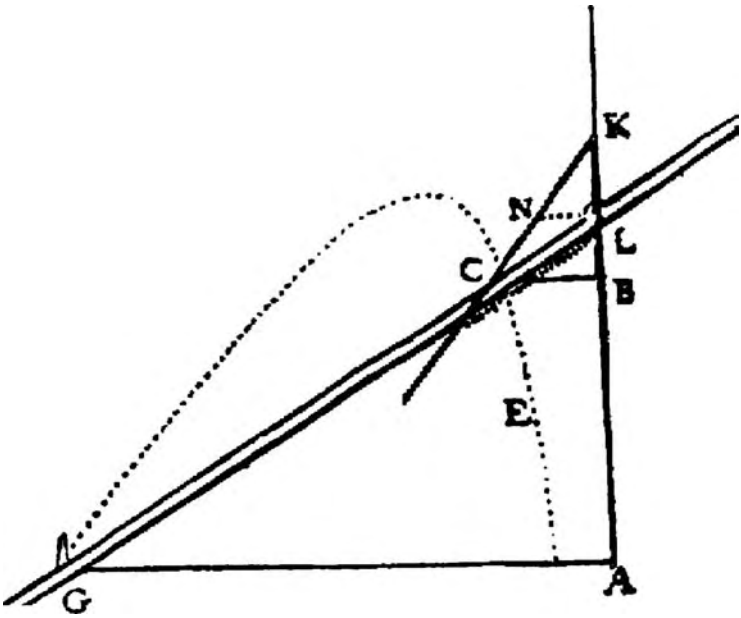


Figure 3. Ruler-and-slide.

In modern terminology, the 'four lines' problem can be stated as follows. Let k be a positive real number, D_1, D_2, D_3, D_4 four lines in the plane, and α_1, α_2, α_3, α_4 four angular magnitudes. Through a point C in the plane draw four lines Δ_1, Δ_2, Δ_3, Δ_4, where Δ_i meets D_i, at a point H_i say, in the angle α_i. Then it is required to find the locus of the points C (Figure 4) such that

$$CH_1CH_2 = kCH_3CH_4. \tag{2}$$

This can be generalized to 'the problem of Pappus on $2n$ lines' as follows. Given two sets of n lines in the plane, two sets of angular magnitudes, and a number k, what is the locus of the points in the plane the ratio of the products of whose distances from one set of lines to those from the other, all measured at the given angles, is equal to k? In the corresponding problem on $2n-1$ lines, one compares the product of the distances from n of them with that from the other $n-1$ except when $n=3$, when one compares the product of two of the distances with the square of the third. Putting in modern notation, it is

$$\omega = (a_1, b_1, c_1, \ldots, a_n, b_n, c_n, \ldots, a_{2n}, b_{2n}, c_{2n}, k) \in \mathbb{R}^{6n+1}; \tag{3}$$

the case of $2n$ lines leads to the equation

$$\prod_{i=1}^{i=n}(a_i x + b_i y + c_i) - k \prod_{i=n+1}^{i=2n}(a_i x + b_i y + c_i) = 0. \tag{4}$$

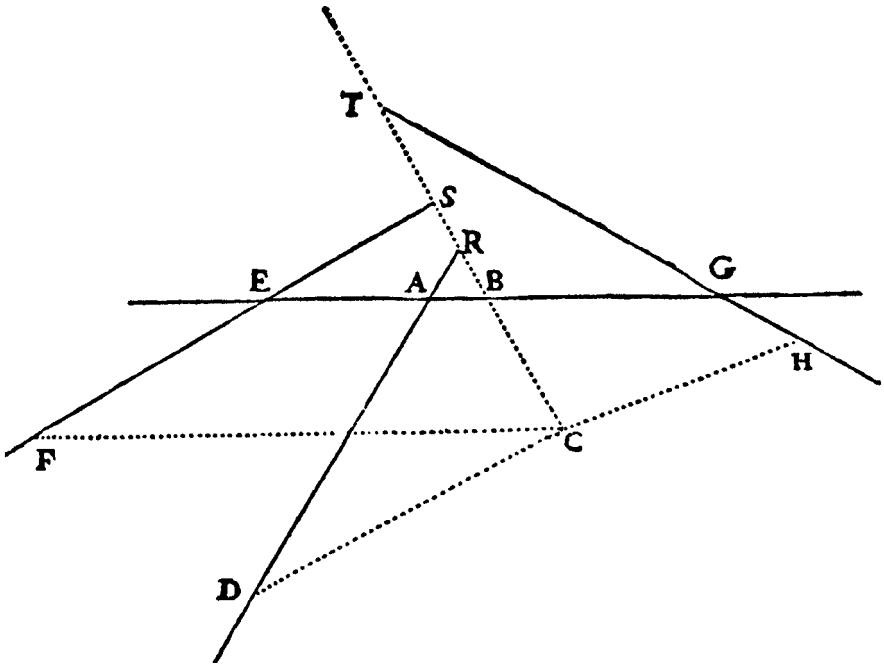


Figure 4. Pappian lines problem. The points B, D, F, H in the figure are the H_i above.

The difficulty of a general solution without analytical tools is evident. The solution being lost, the problem was a test-case of the first importance for Descartes. Claude Hardy (1598–1678), a contemporary at the time of its solution, later reported to Leibniz the difficulties that Descartes had met in solving it (it took him six weeks), which 'disabused him of the small opinion he had held of the analysis of the ancients'.

In the *Géométrie*, the problem became a model for the generation of acceptable curves according to the new criterion, *algebraicity*. It appears in Book I, following preliminaries about expressing in an equation the 'distance' from a point to a line given as an affine function of the coordinates, $CH = ax + by + c$. This requires the use of algebraic symbolism à la Vieta and (implicitly) of a coordinate system. Descartes uses (without saying so) a non-orthogonal frame of reference without emphasizing the use of coordinates (x and y are lengths) and employs similar triangles instead of calculations with distances. The problem reappears in Book II when expressing by an equation the locus of three or four lines, in which he effectively forms products of 'distances'. He finds an equation of the second degree for one of the variables,

$$y(a_1x + b_1y + c_1) = k(a_2x + b_2y + c_2)(a_3x + b_3y + c_3), \tag{5}$$

or, in modern notation,

$$x^2P_0(y) + xP_1(y) + P_2(y) = 0, \tag{6}$$

where the Ps are polynomials in y; this is the first case in history of a geometric locus being expressed by an equation. The formulae for the solution contain, as well as the 'four operations', only square roots, which are constructible.

Descartes identifies point by point the solution as a hyperbola with the equation

$$yy = cy - (c/b)xy + ay - ac \tag{7}$$

by a ruler-and-compass construction of the roots of his equation, without ever recognizing equations of the first degree as those of straight lines. For five lines, one has the equation

$$y(a_1x + b_1y + c_1)(a_2x + b_2y + c_2) = k(a_3x + b_3y + c_3)(a_4x + b_4y + c_4) \tag{8}$$

of degree 3 in y and of degree 2 in x, which again requires the construction for each fixed y the solution of a quadratic equation. For $2n$ lines, one has

$$x^k P_0(y) + x^{k-1} P_1(y) + \cdots + P_k(y) = 0, \tag{9}$$

of which Descartes tried to construct solutions in x for each fixed y. Thus, for a given y, the construction of an *arbitrary* point of the curve is the key step, summed up in the Leitmotif 'all the points [of a geometrical curve] must somehow be related to all the points in a straight line'.

Descartes thus recognized that the equation $F(x, y) = 0$ determines a locus and thus a curve, which arises from the excess of unknowns (two in the case of a single equation); this is the basic idea behind the plane loci that Fermat had described more clearly in the *Isagoge* (1629). It is essentially analytic in the Cartesian sense; the equating of products as above, which is the basis of the strategy of loci, only legitimizes the existence of a point C of the required kind, that is, something that has been taken for granted. 'Let us suppose the thing has been done' is an expression Descartes uses again and again. Thus, when applied to loci, the Cartesian geometry 'of coordinates' is also 'analytic' in a natural way.

For Descartes, the Pappus curves served as models; for they could generate, in accordance with a complexity measured by $2n$ and also in accordance with the second criterion, algebraicity, all the curves that he would henceforth regard as acceptable. He noticed, however, that this classification is inadequate as to complexity. He also obtained Pappus curves for five, six and ten lines, still constructible by ruler and compass, such as the CP, as the curve for five lines with four parallel and equidistant and the fifth meeting them orthogonally: the complexity does indeed depend as much on the disposition of the lines as on their number. He therefore abandoned this criterion in favour of the more practical one of intrinsic complexity–simplicity, that is, the degree of the equation, or rather its 'genus' (he continued to group degrees together in pairs $(2n - 1, 2n)$). The quadratrix and spiral were again rejected by this criterion, since it was not possible to construct an arbitrary point: for the quadratrix, only points with certain rational abscissae, such as $k/2^n$, could be constructed.

'Simplicity' thus became a dimension in the classification of acceptable curves: to breach it would be to commit 'a mistake'. Descartes even conjectured that every algebraic curve (in the modern sense) is a Pappus curve, a false assertion (an arbitrary curve has

'too many' coefficients) that was immediately criticized by his successors, such as Newton [Bos, 1981].

Following Descartes, the supremacy of algebraic criteria became established: curves were defined by equations with integer degrees. Algebra thus brought to geometry the most natural hierarchies and principles of classification. This was extended by Newton to fractional and irrational exponents, and by Leibniz to 'variable' exponents (*gradus indefinitus*, or transcendence in the sense of Leibniz).

7 TANGENTS

Descartes continues his promotion of the algebraic method in Book II, emphasizing the importance of determining the angle made by two curves at one of their points of intersection as 'the most useful and general [question ...] I have ever wanted to answer'. This angle is that between the tangents, hence also between the normals. He therefore wished to construct the normal at a point C of an algebraic curve by finding a point P on it. He wrote that, at the intersection with the curve of the circle with centre P and containing C, there are two points coincident with C, so that the algebraic equation for their common ordinate has a fixed double root, the ordinate of C in Figure 5.

The method required knowledge of how to eliminate one variable between two algebraic equations (a point also considered by Fermat), and then to state the condition for which one knows of an equation that a root is a double root. Descartes did this either by identifying the coefficient of $(y-e)^2$ in (E) (when the curve is an ellipse for example) or, generally, by dividing (E) by $(y-e)^2$. In the case of the CP, where (E) is of degree six, the division of the first (complex) member by $(y-e)^2$ is done by identifying the quotient via four indeterminate coefficients. If these techniques of identification were completely new at the time, the method of indeterminate coefficients later became universal, from Leibniz and Newton to the present day. Profoundly algebraic, the Cartesian method of tangents introduced the concept of a double point. Not being very convenient, it would hardly survive into posterity; following Fermat, tangents were determined by Newton in the 1660s and Leibniz in the 1670s using infinitesimal and differential methods (§5, §4). After 1700,

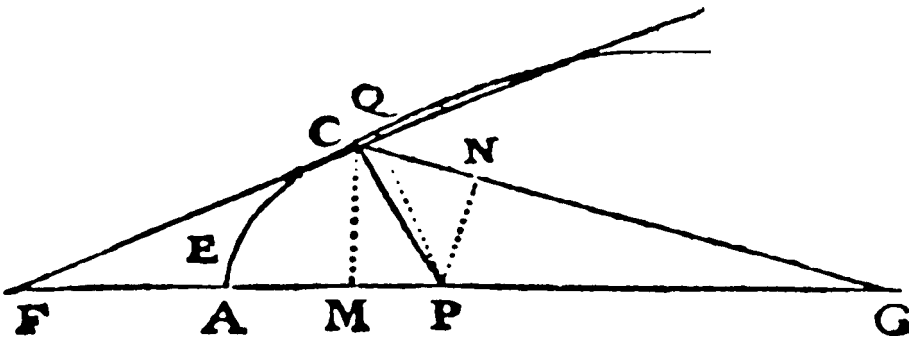


Figure 5. Construction of the normal.

the many treatises on curves combined Cartesian algebraic concepts (equations, multiple points) with infinitesimal local considerations (cusps, points of inflection).

8 OVALS

One section, 'Ovals', occupies some fifteen pages of Book II. Descartes defines step by step the construction of 'four new kinds of oval of use in optics' according to the values of the parameters (lengths of segments), of which the modern reader will recognize no more than two (compare Tannery's 'clarification' in [ATx, 325–328]). Descartes, who shares with Willebrord Snell the discovery of the laws of refraction, introduces these curves in the context of the *Dioptrique*: to determine the surfaces of glass such that after refraction the rays meet at the same point.

The rhetorical value of the discussion of ovals is important for the Descartes of the *DM*: they link a work on physics (the *Dioptrique*) with one on mathematics (the *Géométrie*). Van Schooten notes that they are acceptable in the algebraic sense: in modern terms, they are curves of degree four (their bifocal definition is simpler: $FM \pm \lambda F'M = c$, where λ is a positive real number). Their optical use for reflection and refraction then requires the determination of tangents, justifying the above calculation of normals. Like the compass curves, ovals also admit, for certain values of parameters, a mechanical construction, depicted by Descartes as a taut string. Finally, like the CP, they present another generalization of the central conics, the ellipse and hyperbola (a point repeatedly emphasized by Descartes), in the sense that for certain values of the parameters the ovals 'degenerate' into these conics.

9 ALGEBRAIC EQUATIONS

Descartes motivates this study, which occupies the beginning of Book III, with the desire of avoiding one of the two opposing 'mistakes' when solving a problem that he denounces; namely, the use of methods that are either excessive in relation to their object and thus superfluous, or inadequate and thus unsuccessful. In the mathematical context of the *Géométrie*, the 'clearly conceived' of the *DM* thus becomes the 'most simple'. He essentially attaches to each problem a certain level of complexity and an appropriate methodology. In algebra, he applies this principle to the reduction of algebraic equations, often assumed implicitly to have rational coefficients.

The abundance and variety of results in this section is remarkable. A number of the interesting results presented are not altogether new, some being due to Girolamo Cardano, Thomas Harriot and Albert Girard. The exposition is, however, clear and systematic, and expressed for the first time in history in modern notation; one finds it again in Jan Hudde's *De reductione equationum* (published in G59). These results were taken up and extended by Newton in *Arithmetica universalis* [1707], in lectures between 1673 and 1683.

For the sake of simplicity, Descartes wanted to consider only irreducible polynomials (in $Z[X]$), for otherwise they would be further simplified, and he therefore studied factorizations. He first states without proof that the maximum number of roots that an equation 'can have' is equal to its 'dimension' (degree), as does Girard in his *Invention nouvelle en algèbre* of 1629. When the total number of 'true' (positive) and 'false' (negative) roots is

less than the dimension, one can, according to Descartes, artificially adjoint 'imaginary' roots, a naïve term coined by him but not described (he does not even write $\sqrt{-2}$, like Girard). He also proves that for a polynomial P to be divisible by $(X - a)$, a 'binomial' containing a 'true' root a, it is necessary and sufficient that $P(a)$ be zero (and likewise with $(X + a)$ and $P(-a)$). He then uses his indeterminate coefficients to describe the division of a polynomial by $(X - a)$. It was important for him to know at least one root. For an equation with rational coefficients, he studies the rational roots by first obtaining integer coefficients using multiplication by a suitable denominator; the possible integer roots are then among the divisors of the constant term. This method is also to be found in Girard.

Descartes is also interested in the number of real roots, and asserts without justification that the *maximum* number of positive *or* negative roots of an equation is that of the alternances *or* permanences of the signs '$+$' and '$-$' between consecutive coefficients. This is the celebrated 'rule of signs', which earned unfounded criticism for Descartes [Montucla, 1799, vol. 2, 114–115]. Newton took up and extended the matter in the *De limitibus aequationum*, which concludes the *AU*. The result was proved in the 18th century, in particular by J. De Gua and J.A. Segner, and led to the theorem of Jacques Sturm (1829) on the number of real roots contained in a given interval $[a, b]$.

Descartes also transformed equations by various mappings, such as $x \to x - a$ and $x \to ax$, especially to annihilate the 'second term', giving $x^4 + px^2 + qx + r$ in the case of degree four. He can render all the roots as true or as false, by a method described in Harriot's posthumous work *Artis analyticae praxis* of 1631. The nature, positive or negative, of a root is thus immaterial to Descartes. Whereas Vieta categorically rejected negative numbers, Descartes, like Girard, accepted them in algebra as roots of equations but rejected them in geometry, prefixing the letters with '$\pm$'. By a change of variables, he can always fix in advance the value of the constant term of an equation. Without proof he then displays the value of a cubic resolvent for the (almost) general quartic, describing its factorization as a product of two quadratics. Undoubtedly he adapted this method from the *Ars magna* of Cardano and Ferrari, but remarkably he invoked it here to determine the resolvents of certain examples, some of them parametric:

$$z^4 * + \left(\frac{1}{2}a^2 - c^2\right)z^2 - \left(a^3 + ac^2\right)z + \left(\frac{5}{16}a^4 - \frac{1}{4}a^2c^2\right) = 0. \tag{10}$$

(In the *Géométrie*, '$*$' denotes the position of a 'missing' term and '.' denotes $\pm$.) Degree four is thus reduced algebraically to degree three; it is further reduced geometrically, using the circle and parabola. Thus everything converges towards degree three, and thus one must know how to construct the solutions.

10 THE 'CONSTRUCTION' OF EQUATIONS

The need to solve algebraic equations stems from the problem of Pappus and the scheme of 'loci'. The *geometrical* construction of solutions, which is the third thread of Ariadne in the *Géométrie*, was first and foremost a result of the Cartesian theory of knowledge. For Descartes, to know was in fact to construct. In spite of algebraic appearances, he never departed from a 'Greek' constructivist position. In accordance with the mood of the times,

this meant *construction by curves*. On this point, which ends Book III, Descartes is close to Fermat but differs from Vieta, who also proposed auxiliary curves but confined himself to construction for *determined* problems (like those of the 'expressions' in section 3 above) and not for 'loci'.

Descartes begins with the 'solid problems' naturally identified with the construction of solutions of irreducible equations of the third and fourth degrees. Invoking his 'simplicity', he restricted himself to using in the construction only the circle and parabola. Using letters to denote positive real parameters, or lengths of intervals, the list of equations reduces to these two:

$$z^3 = *pz\,.\,q \quad \text{and} \quad z^4 = *pzz\,.\,qz\,.\,r. \tag{11}$$

Descartes describes his two basic constructions, by 'circle-vertex' and parabola for degree three (basic), and 'circle-shift' and parabola for degree four (Figure 6). The positive roots of the equations correspond to the lengths of the projections on the axis *CDK* of the points of intersection of the two curves lying to the left of the axis, and the negative roots to those on the other side [Serfati, 2003]: the 'modern' expression as an equation is irreproachable. Then, as applications, come the Greek problems of means and trisection. To

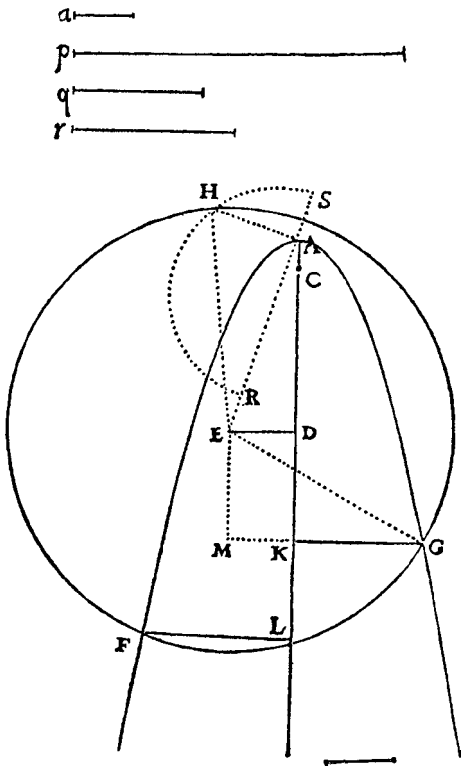


Figure 6. Construction of an equation.

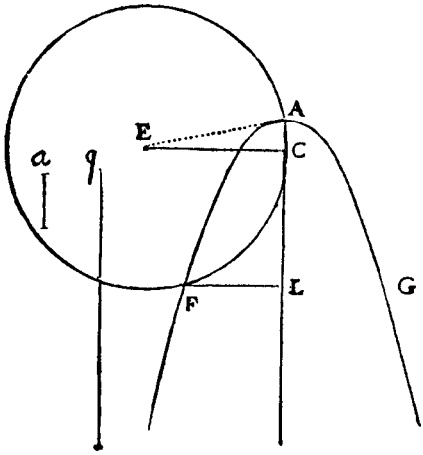


Figure 7. Problem of proportional means.

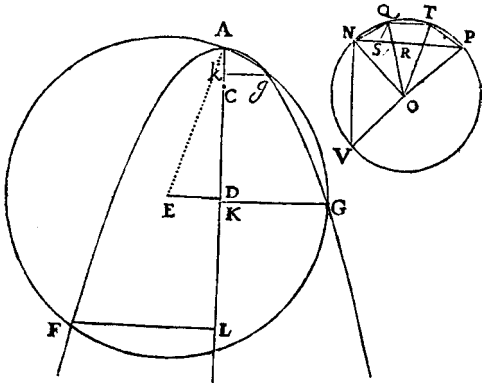


Figure 8. Construction for a cubic equation.

find two proportional means between a and q, that is, to solve $\frac{a}{z} = \frac{z}{y} = \frac{y}{q}$, is to construct the solutions of $z^3 = **aaq$ by the basic circle-vertex method (Figure 7).

The other example, that of trisecting an angle, is equivalent to solving $z^3 = *3z - q$: Descartes proves this using the similarity of isosceles triangles with one angle equal. He then uses the basic construction to find the solutions of this cubic (Figure 8).

These two Greek questions embody for him the essence of solid problems: he shows in effect, like Vieta in the *Supplementum geometriae* (1593), that to solve *any* incomplete equation of degree three of the form $z^3 = * + pz + q$ or $z^3 = * + pz - q$, comes down to trisecting a certain angle. Like Vieta, he concludes without a convincing proof that problems such as trisection that lead to irreducible cubic equations cannot be solved by ruler and compasses, a cornerstone in the characterization of constructible numbers by Pierre Wantzel in 1837. Descartes then describes the problems to be tackled in the future, simple

extensions of the constituent problems (means and trisection) of a higher order, asserting that they 'cannot be constructed from any of the conic sections'.

In all these cases, Descartes was concerned to construct solutions *using curves*. These latter, which we may call 'constructing curves', and they were thus means and not ends [Bos, 1981]. And the constructing curves involved became more and more complicated (like the CP): Descartes never drew a single one! In modern terms, the method regarded an algebraic equation $H(x) = 0$ as the *resultant* of eliminating y between $F(x, y) = 0$ and $G(x, y) = 0$. To construct the solutions of $H(x) = 0$, it suffices to make a suitable choice of F and G, and then to study graphically the abscissae of the points of intersection of the constructing curves with equations $F = 0$ and $G = 0$, the skill of the geometer lying in the 'most simple' choice of F and G, which is clearly an ill-defined concept.

Descartes asserted falsely that to construct the solutions of an equation of degree n, it is necessary to cut a circle (in all cases) with a curve of degree $(n - 1)$, an error that was pointed out by Fermat and Jakob Bernoulli, among others. Forgotten today, these problems of 'construction from equations' survived for a long time [Bos, 1984]. In the *Equationum constructio linearis*, an appendix to *AU*, Newton rightly criticized as ill defined this Cartesian notion of simplicity, contesting the systematic use of the circle, the parabola being for him, *because of his equation*, the simplest [Montucla, 1799, vol. 2, 128]. But the vast array of potential constructing curves served another purpose in the *Géométrie*: the construction of a hierarchical universe of new curves, extending ad infinitum the classification of the ancients.

11 FROM THE *GÉOMÉTRIE* TO THE *ENUMERATIO*

Immediately after the *Géométrie* was published, the question of tangents became the subject of a quarrel, stirred up by Fermat into a famous dispute (see [ATi, ii] for the correspondence between Descartes, Mersenne, Fermat and Roberval at the end of 1637). In a short manuscript, 'Methodus ad disquirendam maximam et minimam' ('The method of maxima and minima'), Fermat described another method, of 'differential' inspiration, that of Descartes being algebraic. Written soon after the *Isagoge*, this was one of a number of unpublished manuscripts circulated by Fermat among his friends in Paris. Descartes, convinced of the complete correctness of his own approach, displayed a sort of blindness [Milhaud, 1921, 149–175; Mahoney, 1994, 170–193]. Roberval naturally supported Fermat, for his part proposing another procedure for tangents, by composition of motions. Descartes had the support of two old friends, Hardy and Claude Mydorge (1585–1647), the author of *De sectionibus conicis* (Paris, 1631, 1639, 1644); also Girard Desargues (1591–1661); and above all De Beaune, a staunch Cartesian, who, in his *Notae* on G49, became a promoter of the new analysis.

After 1649, the text became a long-lasting object of study for European mathematicians and a veritable bedside read for geometers, while the faithful Latin translation by the disciple van Schooten ensured its wide dissemination. The ample commentaries, painstakingly completed by De Beaune and much longer than the *Géométrie* itself, were indispensable in explaining Descartes's ideas to his contemporaries, clarifying obscurities, reconstructing omitted calculations and also producing new constructions and loci.

Over many decades, a considerable number of treatises, too many to cite here, made essential reference to Descartes and the *Géométrie*. Successive modifications of the concepts took various directions, very progressively at first and often based only on examples. Thus for a long time there was, as with Descartes, only a single 'true' axis and a corresponding asymmetry of coordinates, a term absent from Descartes, the modern nomenclature (coordinates, abscissa, ordinate, constant, parameter) being due to Leibniz (1692). Similarly, the *folium* of Descartes,

$$x^3 + y^3 - axy = 0, \tag{12}$$

survived for a long time (up to Huygens in 1692) in its restricted form as a loop, when its manifest incompleteness served as a stimulus for the introduction of negative coordinates. From 1655, the *De sectionibus conicis* of John Wallis used the Cartesian method for expressing algebraically the ancient geometric definitions and properties of the conics of Apollonius, systematically interpreting x and y as having any sign. *De sectionibus* favoured the modern view of conics as plane curves rather than spatial intersections ('solids'). However, Wallis did accuse Descartes unjustly of having plagiarized Harriot ([Montucla, 1799, vol. 2, 115–120] and compare §2).

The *Lieux géométriques* (1679) of Philippe de La Hire (1640–1718) introduced the symmetrization of the coordinates. Leibniz, who knew the *Géométrie* in Paris no later than 1674, made full use of its methods in his arithmetic quadrature of the circle around 1674, his first discovery. G59 was decisive in directing the interest of the young Newton towards mathematics in the years 1664 and 1665. From his 'Method of fluxions and infinite series' (1671?) onwards, he skillfully manipulated the geometry of coordinates. But in his 'Equationum constructio' he remained ambiguous over Descartes, strangely siding with the 'antique' conservativeness of Barrow against the Cartesian 'modernity'.

After 1637, Descartes turned away from all theoretical work in mathematics, even the new questions he had raised in the *Géométrie* (except under duress, as in the case of Fermat and the tangents). He took no interest in general equations of degree two and still less in the classification of cubics. This was accomplished by Newton in the *Enumeratio linearum tertii ordinis* (1676?, published in 1704), a remarkable natural extension of Descartes's work which listed 72 types of cubic though missing six that were found in the 18th century. Breaking away from the original Cartesian practice, the *Enumeratio* classified cubics in a modern fashion according to the values of the parameters and the representations using two orthogonal axes with the same status and an origin of exchangeable coordinates of arbitrary sign, like the parameters. This was an important step: it is not the same as accepting, like Descartes, negative roots of equations and then representing them in a figure, and supposing that letters stand for quantities, unknown or known, of any sign.

12 '*PROLES SINE MATRE CREATA*'

In his *Géométrie*, Descartes organized a mathematical revolution by establishing, in a polished, effective and clear manner, a relation between curves and algebraic calculation, both the continuum of geometry and the discontinuity of number. Today, the use of coordinates in visualizing a curve by means of its equation is an almost automatic process. However, the text is not clear for two reasons: on the one hand by the conflict of criteria, on the other

hand by the Cartesian strategy of secrecy. These reasons can explain its obscurity. In a book where Descartes is supposed to be describing the principles of his method, the reader does not get to the heart of the matter: the correspondence between curves and equations is not made explicit by the use of coordinates. It is written like a self-teaching manual: 'the *Géométrie* does not deliver to the reader the considerable reform of which it is the fruit, but only a set of directions for use' [Costabel, 1987, 220].

Descartes actually proposes two criteria of acceptability for curves that are not logically connected. The first is the CM criterion, exemplified by the compasses and ruler-and-slide. This was his first programme, spontaneous, inherited from his youth, and not entirely renounced at the time of writing the *Géométrie*. The second is the algebraic criterion, which concerns the possibility of constructing point by point the roots of equations derived from Pappus curves. This is, in contrast, a reasoned criterion, dictated by the exigencies of simplicity and effectiveness encountered while writing the final draft. To ensure universal acceptability, Descartes hoped that the two criteria were logically equivalent, that is, every curve with an equation has the CM property, and vice versa. He then postulated their equivalence, although this was not established until the 19th century [Kempe, 1876]. He also hoped to 'prove' it using the ubiquity of the CP, and obtained from each of the criteria. In the *Géométrie* itself, the conflict between the criteria turns in favour of the algebraic, admittedly not without trouble but clearly enough: 'and in some other way imagining the description of a curved line, provided it is of the type that I call geometrical, it is always possible to find an equation that determines all its points in this way' (pp. $395 \equiv 27$). The conflict nevertheless has an effect 'in real time', adding to the difficulty of the text in the eye of the reader.

Posterity has clearly retained the algebraic criterion for classifying curves, being both simple and practical, subsequently improved (in 1695) by replacing 'genus' by 'degree'. Thus the criterion of algebraicity that enables one to envisage 'all the (geometric) curves' has the potential to extend the class of acceptable curves in a fashion inconceivable to the Ancients, who knew only a small number of curves occurring individually 'on the ground'.

On receiving the *Géométrie*, Fermat sent to Descartes in return (via Pierre de Carcavi and Mersenne) two short manuscripts, the treatise 'De maximis et minimis' mentioned earlier and an 'Introduction to plane and solid loci' ('*Ad locos planos et solidos isagoge*'). The *Isagoge*, a short treatise on analytical geometry written around 1629 in which frequent reference is made to the *De emendatione* of Vieta, presents more explicitly than Descartes the scheme of 'loci' ('planes' and even 'solids'). The dispute for priority as the true discoverer of analytical geometry, Descartes or Fermat, which consumed enormous quantities of ink [Milhaud, 1921; Mahoney, 1994], was nevertheless to no purpose. They had both begun at the point where Vieta had left off, and had discovered, independently and at almost the same time, two theories on the same subject but with non-comparable extensions, processes, objectives and notation. A detailed analysis shows the supremacy of one or other of the two protagonists, depending on the point in question [Boyer, 1956, 74–102; Brunschvicg, 1981, 99–126]. If, for example, Fermat gave precedence to the scheme of 'loci', he did not, unlike Descartes, accompany with it any organized view of the hierarchy of curves produced. None of Fermat's manuscripts was published during his lifetime, while the *Essais*, widely disseminated in Europe, aroused a profound interest in the mathematical community. The semi-rhetorical style of Fermat, who was a follower of Vieta, was not

conducive to working out calculations, unlike the Cartesian exponential notation, which was adopted rapidly and almost universally.

The ease that Descartes displays in handling symbolic notation and the familiarity that we have with it today must not obscure its profound novelty at the time, or the 'shock' [Costabel, 1987, 218] of that supposedly 'geometric' work, but with every page covered by calculations, letters and new symbols, provoked among his contemporaries. The use of the exponent, and of a specific sign for equality—while different from that of today (Recorde), it still destroys the syntax of natural language [Serfati, 1998]—just like the systematic literalization of Vieta, were decisive in the advent of the new symbolism. The statement in modern terms of Cardano's formula in one line in Book III proves to the reader the clear superiority of this symbolism over the laborious rhetoric of Cardano. Contemporary mathematicians were not mistaken in using the *Géométrie* as a 'Rosetta stone' for deciphering symbolism.

Conversely, it is the first text in history to be directly accessible to mathematicians of today. By the systematic use of substitutions (his *Art combinatoire*), Leibniz continued the implementation of what was not just a 'change of notation' but a radical modification of modes of mathematical thought [Serfati, 1998]. Fractional and literal exponents were added by Newton.

In 1630, Descartes declared to Mersenne that he was 'tired of mathematics', and he meant it. Leaving for Holland, he thought of the mathematical model of his youth as over and turned to metaphysics. In 1637, however, when faced with the need to find applications of his *DM*, he returned briefly to mathematics. But while he was proud of the results found, he never intended to continue in mathematics. So incontestably the *Géométrie* represents a culmination in his work and not as an avenue opening towards the future. After 1637, he devoted himself exclusively to philosophy, while occasionally studying with his correspondents certain mathematical problems. Some of these lay in areas previously rejected, such as the question of the divisors of an integer (including, strangely enough, the integers equal to twice the sum of their proper divisors), and the study of non-geometric curves like the 'roulette', or cycloid, or again an 'inverse-tangent' problem posed by de Beaune, the first in the history of differential geometry, of which the solutions are transcendental curves. Descartes's correspondence with Mersenne after 1637 shows that he was conscious of having produced an exceptional mathematical work that few of his contemporaries seemed able to understand. Admittedly, he had simply sought to introduce, through his analytical geometry, a method into geometry, the algebra being merely a tool; but in fact he had achieved more, and with his customary pride, he gave a good account of himself. Henceforth, he said to Mersenne at the end of 1638, he had no need to go any further in mathematics.

In a famous phrase, Chasles described the *Géométrie* as 'a child without a parent' ('*proles sine matre creata*'). This judgement needs some slight qualification. It is true that Descartes borrowed little from his predecessors, certainly not from Fermat but perhaps from Girard and Harriot; on the other hand, he had certainly read Pappus, Cardano and above all Vieta, although he only admitted this in Holland after 1628, declaring proudly at the end of December 1637 that he had 'begun where he [Vieta] had left off'. It was nevertheless to a large extent independently of his contemporaries, and guided by a specific 'physical' view of the world, that Descartes wrote the *Géométrie*, a work as exceptional

for its contents (analytic geometry and the analytical method) as for its style (symbolic notation), which progressively organized the final breakaway from the ancient and mediaeval mathematical world, pre-symbolic and 'specious'. Only then could one pass into the age of the 'calculus'. Let us emphasize this obvious fact: Leibniz, while rightly deploring the absence from Descartes's mathematical thought of any infinite operation, a concept that is indeed foreign to the Cartesian system of the world, nevertheless could only criticize Descartes retrospectively.

BIBLIOGRAPHY

Baillet, A. 1691. *La vie de Monsieur Descartes*, Paris: Hortemels. [Repr. Hildesheim and New York: Olms, 1972.]

Bos, H.J.M. 1981. 'On the representation of curves in Descartes's *Geometrie*', *Archive for history of exact sciences, 24*, 295–338.

Bos, H.J.M. 1984. 'Arguments on motivation in the rise and decline of a mathematical theory: the 'Construction of equations', 1637-*ca*. 1750', *Archive for history of exact sciences, 30*, 331–379.

Bos, H.J.M. 1991. 'The structure of Descartes's *Geometrie*', in *Lectures in the history of mathematics*, Providence: American and London Mathematical Societies, 37–57.

Bos, H.J.M. 2001. *Redefining geometrical exactness. Descartes' transformation of the early modern concept of construction*, New York: Springer.

Boyer, C.B. 1956. *History of analytic geometry*, New York: Scripta Mathematica.

Brunschvicg, L. 1981. *Les étapes de la philosophie mathématique*, Paris: Blanchard.

Chasles, M. 1837. *Aperçu historique sur l'origine et le développement des méthodes en géométrie, particulièrement de celles qui se rapportent à la géométrie moderne*, Bruxelles: Hayez. [Repr. Paris: Gauthier–Villars, 1875; Paris: Gabay, 1989.]

Costabel, P. 1987. 'Les *Essais de la Méthode* et la réforme mathématique', in *Le Discours et sa Méthode* (ed. N. Grimaldi and J.-L. Marion), Paris: P.U.F., 213–228.

[AT] Descartes, R. *Oeuvres* (ed. C. Adam and P. Tannery), 13 vols., Paris: Cerf, 1897–1913. [Various reprs., esp. of vols. 1–11, Paris: Vrin, 1964 onwards; paperback ed. 1996. ATi–ATv contains Descartes's correspondence, notably the mathematical. ATvi contains *DM* and the *Essais*, and a facsimile of G37, with original pagination indicated. References to the commentaries of van Schooten in G49 et G59 are indicated by capital letters. ATx contains the correspondence with Beeckman, the *Olympica*, the *Regulae*, the *Excerpta mathematica* (Mathematical fragments), *De solidorum elementis*, and *Calcul de M. Descartes—Introduction à sa Géométrie*.]

Descartes, R. 2002. *Règles pour la direction de l'esprit* (trans. J. Brunschwicg), Paris: Livre de Poche.

Kempe, A.B. 1876. 'On a general method of describing plane curves of the nth degree by linkwork', *Proceedings of the London Mathematical Society*, *7*, 213–216.

Mahoney, M.S. 1994. *The mathematical career of Pierre de Fermat, 1601–1665*, Princeton: Princeton University Press.

Milhaud, G. 1921. *Descartes savant*, Paris: Alcan.

Montucla, J.-F. 1799–1802. *Histoire des mathématiques*, 2nd ed., 4 vols., Paris: Agasse. [Repr. Paris: Blanchard, 1960. See §21.]

Newton, I. 1707. *Arithmetica universalis* (ed. W. Whiston), Cambridge: Typis Academicis. [Various later eds. Lectures dating from around 1673.]

Rodis-Lewis, G. 1995. *Biographie de Descartes*, Paris: Calmann-Lévy.

Serfati, M. 1993. 'Les compas cartésiens', *Archives de philosophie, 56*, 197–230.

Serfati, M. 1994. '*Regulae* et mathématiques', *Theoria* (San-Sebastiàn), *(2) 11*, 61–108.

Serfati, M. 1998. ‘Descartes et la constitution de l’écriture symbolique mathématique’, in his (ed.), *Pour Descartes*, *Revue d’Histoire des Sciences*, *51*, nos. 2/3, 237–289.

Serfati, M. 2003. ‘Le développement de la pensée mathématique du jeune Descartes’, in his (ed.), *De la methode*, Besançon: Presses Universitaires Franc-Comtoises, 39–104.

Struik, D.J. 1969. *A source book in mathematics 1200–1800*, Cambridge, MA: Harvard University Press. [Repr. Princeton: Princeton University Press, 1986. Contains some English translations of Descartes and Fermat, and also of Girard, Vieta, Newton and Leibniz.]

Whiteside, D.T. 1961. ‘Patterns of mathematical thought in the later seventeenth century’, *Archive for history of exact sciences*, *1*, 179–388.

CHAPTER 2

JOHN WALLIS, *ARITHMETICA INFINITORUM* (1656)

Jacqueline Stedall

The *Arithmetica infinitorum* was a key text in the 17th-century transition from geometry to algebra and in the development of infinite series and the integral calculus.

First publication. In *Operum mathematicorum*, vol. 2, Oxford: Oxford University Press, 1656, 1–199. [Digital copy available in the database 'Early English books online' (EEBO).]

Reprint. In *Opera mathematica*, vol. 1, Oxford: Oxford University Press, 1695, 355–478. [Edition photorepr. Hildesheim: Olms, 1972.]

English translation. *The arithmetic of infinitesimals* (trans. J.A. Stedall), New York: Springer-Verlag, 2004.

Related articles: Descartes (§1), Newton (§5), Berkeley (§8), Maclaurin (§10), Euler *Introductio* (§13).

1 BACKGROUND TO THE *ARITHMETICA INFINITORUM*

John Wallis (1616–1703) was Savilian Professor of Geometry at Oxford for over half a century, from 1649 until his death in 1703 at the age of 87. The professorship was his reward for his services as a code-breaker to the Parliamentarians during the civil wars of the 1640s. Wallis came to Oxford with no more than a little self-taught mathematics, but within a few years produced his most important work, the *Arithmetica infinitorum*. The book was begun in 1651 and completed in early 1655 and was printed by July of that year, but it was published only in 1656 in a compilation of Wallis's mathematical works, the *Operum mathematicorum* (1656–1657). It was reprinted with minor alterations towards the end of Wallis's life, in 1695, in a much larger set of collected works, the *Opera mathematica* (1693–1699).

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

Table 1. Contents of Wallis's book.

Propositions	Page	Contents
1–52	1	Sums of integer powers and geometric corollaries.
53–85	42	Sums of fractional powers and geometric corollaries.
86–102	67	Sums of negative powers and geometric corollaries.
103–168	77	Sums of compound quantities and geometric corollaries.
169–186	136	Properties of figurate numbers.
187–191	167	Interpolation of tables.
'Idem aliter'	181	Brouncker's continued fractions.
192–194	193	Interpolations demonstrated geometrically. [End 198.]

The *Arithmetica infinitorum* is not written in chapters or sections, but as series of 194 Propositions (lemmas, theorems, corollaries), some followed by a 'Scholium' or commentary. At the end of the book, in an 'Index propositionum' (p. 198), Wallis himself listed the propositions and their contents. The propositions and their subject matter are grouped in Table 1.

The book was Wallis's masterpiece. It contains the infinite fraction for $4/\pi$ that is now his chief claim to fame; but for his contemporaries the most significant feature was the introduction of new methods, new concepts and new vocabulary. The *Arithmetica infinitorum* stands both chronologically and mathematically at the mid point of the 17th century, drawing together the best ideas from the first half of the century, the algebraic geometry of René Descartes, and the theory of indivisibles of Bonaventura Cavalieri (1598–1647), and preparing the ground for some of the astonishing advances of the second half: the discovery of the general binomial theorem, applications of infinite series, and the integral calculus.

In *De sectionibus conicis*, a book written at the same time as the *Arithmetica infinitorum* and published alongside it [Wallis, 1656a]. Wallis gave the first fully algebraic treatment of conic sections, showing that they could be adequately defined by their equations alone. Thereafter he rarely used the formulae he had so carefully derived, but the process of finding them must have done much to consolidate in his own mind the possibility of exchanging geometric ways of thinking for the arithmetic or algebraic.

A more fundamental source of inspiration in Wallis's mind was the theory of indivisibles, first proposed by Cavalieri in his *Geometria indivisibilibus continuorum nova quadam ratione promota* of 1635 [Andersen, 1985]. Cavalieri based his ideas on the notion of a plane passing at right angles through a given figure and intersecting it in 'all the lines' of the figure. From this there followed his fundamental theorem: that two plane figures could be said to be in the ratio of 'all the lines' of one to 'all the lines' of the other. Later in the *Geometria*, and also in his *Exercitationes geometricae sex* of 1647, Cavalieri tried to avoid the problems of handling infinitely many lines by comparing pairs of lines, but most of his work relied on the comparison of collections of lines, and he successfully applied his methods to a variety of plane figures and solids.

Cavalieri's careful attempts to put his theory on a sound footing were of little concern to Evangelista Torricelli (1608–1647) when he took up some of the same ideas in his *Opera*

geometrica of 1644. For Torricelli a plane figure was simply the sum of its lines, and a solid was a sum of planes or surfaces. It was through Torricelli's *Opera*, more easily available than Cavalieri's *Geometria*, that Wallis and others learned of the theory of indivisibles, but only in the cruder form into which Torricelli had transposed it. Wallis was even more lax than Torricelli: for him a plane figure could be regarded, when convenient, as the sum of its lines, but at other times as a sum of infinitely thin parallelograms. In Proposition 1 of *De sectionibus conicis* he wrote:

> I suppose, as a starting point (according to Bonaventura Cavalieri's geometry of indivisibles) that any plane is constituted, as it were, from an infinite number of parallel lines. Or rather (which I prefer) from an infinite number of parallelograms of equal altitude, the altitude of each of which indeed may be $\frac{1}{\infty}$ of the whole altitude, or an infinitely small part (for let ∞ denote an infinite number), and therefore the altitude of all taken together is equal to the altitude of the figure.

In the *Arithmetica infinitorum* Wallis never discussed the distinction between lines and parallelograms, and his evasion of such fundamental definitions was to draw criticism later. But Wallis was not too concerned because he recognized that, without being too careful about precise meanings, the summation of lines, or 'indivisibles', gave him useful ways of handling the quadrature and cubature of a multitude of curved shapes.

2 METHODS AND RESULTS IN THE *ARITHMETICA INFINITORUM*

Cavalieri and Torricelli had also used indivisibles for the purpose of quadrature, and so had the French Jesuit Grégoire St. Vincent in his massive *Opus geometricum* of 1647, though the latter was unknown to Wallis when he began his work. Wallis's advance over his predecessors was his conversion of the problem of quadrature from geometry to arithmetic. In Figure 1, to find the ratio of the concave area *ATO* defined by the parabola *AO* (with vertex *A*), to the rectangle *ATOD*, Wallis needed to sum the lines *TO*. Since each *TO* equals $(OD)^2$ for the corresponding *OD*, the problem reduced to finding the ratio

$$\frac{0^2 + a^2 + (2a)^2 + (3a)^2 + \cdots + (na)^2}{(na)^2 + (na)^2 + (na)^2 + (na)^2 + \cdots + (na)^2}, \tag{1}$$

where a is the (small) distance between each of the lines *TO*. Thus Wallis needed to find

$$\frac{0^2 + 1^2 + 2^2 + 3^2 + \cdots + n^2}{n^2 + n^2 + n^2 + n^2 + \cdots + n^2} \tag{2}$$

for large values of n (but small values of a since na was fixed and finite). This he did by what he called 'induction' (see Propositions 1, 19 and 39):

$$\frac{0+1=1}{1+1=2} = \frac{3}{6} = \frac{1}{3} + \frac{1}{6}, \qquad \frac{0+1+4=5}{4+4+4=12} = \frac{1}{3} + \frac{1}{12}, \tag{3}$$

$$\frac{0+1+4+9=14}{9+9+9+9=36} = \frac{7}{18} = \frac{1}{3} + \frac{1}{18}, \tag{4}$$

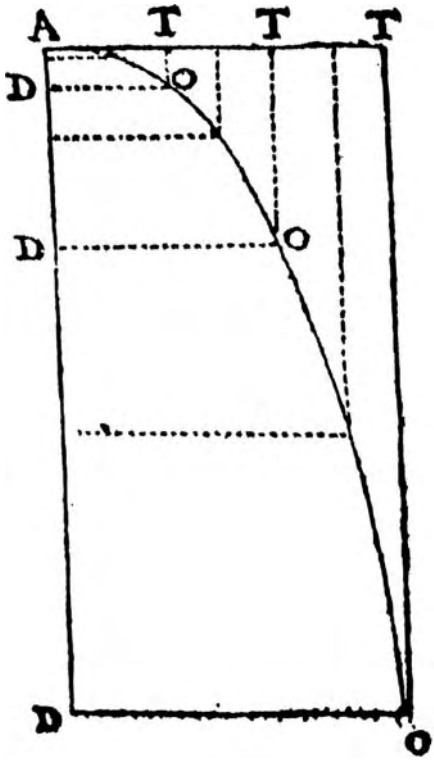


Figure 1. Wallis's analysis of the parabola.

$$\frac{0+1+4+9+16=30}{16+16+16+16+16=80}=\frac{3}{8}=\frac{9}{24}=\frac{1}{3}+\frac{1}{24},\ \cdots. \tag{5}$$

The answer in each case is $\frac{1}{3}$ plus a fraction that becomes smaller as more terms are taken. In fact, Wallis noted, the additional fraction eventually becomes less than any 'assignable quantity' and therefore, if the process is continued infinitely, it may be considered to be zero (see Propositions 20 and 40).

By allowing n to increase and a to decrease Wallis had thus shown that the area *ATO* was one third of *ATOD*, a result already well known to be correct, but he had arrived at it by arithmetic rather than geometry. Thus, just as Cavalieri's method could be described as the geometry of indivisibles, or *geometria indivisibilium*, Wallis's could be described as the arithmetic of infinitesimals, or *arithmetica infinitorum*, and hence the title of his book. Wallis himself translated *arithmetica infinitorum* as 'The Arithmetick of Infinites', but since the 'infinites' in question were in fact infinitely small quantities, 'the arithmetic of infinitesimals' is perhaps a more accurate modern translation.

The above and similar examples were enough to convince Wallis that the method worked, and he set out to extend it by investigating sums of cubes and higher powers,

rapidly arriving at the result that for large n (bearing in mind that na is fixed and finite):

$$\frac{\sum_{k=0}^{n}(ka)^r}{\sum_{k=0}^{n}(na)^r} \approx \frac{1}{r+1}, \tag{6}$$

(here using the modern summation symbol). Wallis introduced the term 'index' to denote the power r of each term, and his next step was to extend the above result by analogy to sequences with fractional or negative index. Again, he continually checked his methods by confirming well known results. In particular he could now perform term-by-term multiplication or division of series, which he interpreted geometrically as multiplication of lines by lines, or division of planes by lines. Here he came close to Grégoire St. Vincent who had performed many similar 'multiplications', but some of Wallis's examples stretch the geometrical imagination to the limit as the potential of his method begins to go beyond anything he could sensibly describe in the language of geometry.

3 THE MOTIVATION TO 'WALLIS'S PRODUCT'

Two final problems of geometry, however, were outstanding: the quadrature of the hyperbola and of the circle, and it was the latter that exercised Wallis for the entire second half of the *Arithmetica infinitorum*. It was easy for him to see that for the area of a quadrant of radius R he needed to sum terms of the form $(R^2 - (ka)^2)^{1/2}$ for $k = 0, 1, 2, \ldots, R/a$, but without the general binomial theorem he had no way of handling such quantities. He could, however, sum terms of the form $(R^{1/p} - (ka)^{1/p})^q$ when p and q were integers, and so he set out to find sums for intermediate, non integer, values by numerical interpolation. In particular he needed the sum for $p = q = 1/2$, which would give the ratio of a square to the inscribed quadrant (in modern notation $4/\pi$), a ratio that Wallis denoted by the symbol $\square$. Wallis's interpolative methods have been fully described elsewhere, and here it need only be said that they required much of the patience and tenacity he must previously have applied to code-breaking [Nunn, 1910–1911; Scott, 1938, 26–64; Stedall, 2002, 159–165].

Wallis himself explained every step in full, often in laborious detail, so that the reader has the experience of entering his mind and sharing at first hand his sense of achievement or frustration as he slowly proceeded towards his goal [Whiteside, 1961]. Wallis began to have a very clear sense that the ratio he was seeking was unlike any other number in mathematics, neither a rational nor a surd, but what we would now describe as 'transcendental' (see [Stedall, 2002, 159–165; Panza, 2004]). Eventually he arrived at the sequence: $\square/2$, 1, $\square$, $\frac{3}{2}$, $4\square/3$, $\frac{15}{8}$, $8\square/5$, $\frac{35}{16}$, ... and realised that although the terms themselves are increasing, the ratios of each term to the previous one, that is, $2/\square$, $\square$, $3/2\square$, $8\square/9$, ... decrease towards 1. Hence he was able to find sequences of upper and lower bounds for $\square$, and so arrived at his famous formula, given here as he wrote it himself (in Proposition 191):

$$\square = \frac{3 \times 3 \times 5 \times 5 \times 7 \times 7 \times \&c.}{2 \times 4 \times 4 \times 6 \times 6 \times 8 \times \&c.}. \tag{7}$$

Wallis's formula is easily proved by the methods of modern analysis, but to have arrived at it with the very limited numerical tools at his disposal was a remarkable achievement.

Wallis's fraction is not the end of the *Arithmetica infinitorum*, for he showed his interpolations to William Brouncker ($c.1620-1684$) who came up with an alternative formulation (see the 'Idem aliter' following Proposition 191):

$$\square = 1\cfrac{1}{2\cfrac{9}{2\cfrac{25}{2\cfrac{49}{2\cfrac{81}{2}}}}}\&c. \tag{8}$$

It is impossible to look at Brouncker's fraction even now without a sense of astonishment. It is written in a form previously unknown in English mathematics, yet it suddenly appears in the *Arithmetica infinitorum* fully fledged, without introduction or preamble. It is not obvious even that Brouncker's fraction is equal to Wallis's, let alone how one form might give rise to the other; were it not for the fact that it appears on the printed page, one would be inclined to say that it was impossible for Brouncker to have found it. Brouncker himself could not be persuaded to explain his methods; Wallis tried to do it for him, but was quickly out of his depth. The most important point is that Brouncker's fraction brought in its train a whole sequence of similar fractions:

$$\square = 1 + \cfrac{1}{2 + \cfrac{9}{2 + \cfrac{25}{2 + \cdots}}}, \quad B = 3 + \cfrac{1}{6 + \cfrac{9}{6 + \cfrac{25}{6 + \cdots}}},$$

$$C = 5 + \cfrac{1}{10 + \cfrac{9}{10 + \cfrac{25}{10 + \cdots}}}, \quad \ldots, \tag{9}$$

with the remarkable multiplicative properties that $\square B = 2^2$, $BC = 4^2$, $CD = 6^2, \ldots$. Further, this sequence of fractions enabled Wallis to find the multipliers he had sought in vain for the sequence:

$$\square/2,\ 1,\ \square,\ \frac{3}{2},\ 4\square/3,\ \frac{15}{8},\ 8\square/5, \ldots, \tag{10}$$

for now he could write it (putting $\square/2 = A$) as:

$$A \times \frac{B}{2} \times \frac{C}{4} \times \frac{D}{6} \times \frac{E}{8} \times \frac{F}{10} \times \frac{G}{12} \times \frac{H}{14} \times \&c. \tag{11}$$

The *Arithmetica infinitorum* goes on to give the first general treatment of continued fractions (the name comes from Wallis's description of them as *fractiones continue fractae*, or 'fractions continually broken'), including a recursive formula for evaluating them 'from the top down' written using the first example of subscript notation.

In the final pages of the book Wallis produced diagrams that illustrated his interpolations geometrically, by the construction of what he called a 'smooth curve' (*curva aequabilis*)

between given points, but this added little to what had gone before, and appears to have been something of an afterthought. He insisted, though, that thanks to his previous work such curves were as well defined as those given by equations, and therefore deserved to be called 'geometric' in Descartes's sense of the word (§1.5). Wallis's *curva aequabilis* was also printed separately at Easter 1655 to advertise his forthcoming book and dedicate it to William Oughtred, and in an enthusiastic letter of thanks Oughtred declared that Wallis had 'opened a way into these profoundest mysteries of art, unknown and not thought out by the ancients' [Rigaud, 1841, vol. 1, 87–88]. Oughtred's letter was written in August 1655, too late for inclusion in the *Arithmetica infinitorum*, but Wallis took care to print it at the front of the second edition 40 years later.

4 REACTIONS TO THE *ARITHMETICA INFINITORUM*

The *Arithmetica infinitorum* quickly circulated amongst mathematicians in England and Europe but, despite the glowing words from Oughtred, it was not an immediate success. Fermat was the first to claim that he had found many of Wallis's results already: 'I have read the *Arithmetica infinitorum* of Wallis and I have great regard for its author. Even though the quadrature both of parabolas and infinite hyperbolas was done by me many long years ago' [Brouncker et alii, 1658, letter IV]. Fermat had indeed found many of Wallis's results, and by not dissimilar means (though for integer indices only), but since they were circulated only in private correspondence, he could not blame Wallis for being unaware of them. Wallis argued (and continued to do so years later) that what he had hoped to achieve were not so much new results as new methods of investigation [Wallis, 1685, 305–306]:

> [Fermat] doth wholly mistake the design of that Treatise; which was not so much to shew a Method of Demonstrating things already known [. . .] as to shew a way of *Investigation* or finding out of things yet unknown [. . .] and that therefore I rather deserved thanks, than blame, when I did not only prove to be true what I had found out; but shewed also, how I found it, and how others might (by those Methods) find the like.

Unfortunately Wallis's methods themselves also came into question. Not only Fermat, but Christiaan Huygens in the Netherlands and Thomas Hobbes in England expressed doubts about Wallis's use of 'induction' which, they argued, was not a secure method of proof [Brouncker et alii, 1658, letter XIII; Huygens, *Works*, vol. 1, 458–460; Hobbes, 1656, 46]. Wallis fell back, as he so often did, on precedent, and claimed that induction had been used by Euclid every time he allowed a triangle to stand for any other triangle of the same kind. This was induction in its most general sense: the inference of a general law from particular instances. Wallis's induction in the *Arithmetica infinitorum* was more than this, and much closer (though not yet identical) to the modern principle of mathematical induction: Wallis showed a proposition to be true for integers $0, 1, 2, \ldots, k$, and then argued that if there was no reason to suppose that the pattern would change then the proposition must hold for any positive integer. Here he cited precedents in the work of François Viète and Henry Briggs, who had made similar claims in their work on angle sections.

The second and more serious stumbling block to Wallis's readers was his lack of clarity on the subject of indivisibles. In *De sectionibus conicis* [Wallis, 1656a] had argued that

a parallelogram of infinitely small altitude was scarcely more than a line, but at the same time that a line must be considered 'dilatable', that is, of some non-zero width, so that infinitely multiplied it would reach to the required altitude. Hobbes saw the weaknesses of Wallis's arguments immediately [Hobbes, 1656, 46]:

> 'The triangle consists as it were' ('as it were' is no phrase of a geometrician) 'of an infinite number of straight lines.' Does it so? Then by your own doctrine, which is, that 'lines have no breadth', the altitude of your triangle consisteth of an infinite number of 'no altitudes', that is of an infinite number of nothings, and consequently the area of your triangle has no quantity. If you say that by the parallels you mean infinitely little parallelograms, you are never the better; for if infinitely little, either they are nothing, or if somewhat, yet seeing that no two sides of a triangle are parallel, those parallels cannot be parallelograms.

Wallis's reply, 'I do not mean precisely a line but a parallelogram whose breadth is very small, viz an aliquot part [divisor] of the whole figures altitude' [Wallis, 1656b, 43], added nothing to what he had said already. In truth the problem did not greatly concern him. For Wallis the justification of his methods was that they worked, and in the *Arithmetica infinitorum* he had clearly deduced numerous correct results for parabolas and hyperbolas. His quadrature of the circle too (that is, his formula for $\square$) was upheld by Brouncker's calculation of the number now known as π, in agreement with the known and accepted value as far as the ninth decimal place [Brouncker et alii, 1658, letter V].

Fermat and Hobbes came to maturity in an era when mathematics was still firmly grounded in classical geometry and Archimedean methods of proof, and both were uneasy about Wallis's apparent abandonment of classical methods. Fermat could have been speaking for either of them when he wrote [Brouncker et alii, 1658, letter XLVI; trans. Wallis, 1685, 305]:

> We advise that you would lay aside (for some time at least) the Notes, Symbols, or Analytick Species (now since Vieta's time, in frequent use,) in the construction and demonstration of Geometrick Problems, and perform them in such method as Euclide and Apollonius were wont to do; that the neatness and elegance of Construction and Demonstration, by them so much affected, do not by degrees grow into disuse.

Apart from finding algebraic formulae for triangular figurate numbers, Wallis's use of 'notes and symbols' was sparing, and it was perhaps not so much the algebraic notation in his text that Fermat or Hobbes objected to as the loss of traditional geometry. Huygens was much younger than Fermat and only twenty-seven when he read the *Arithmetica infinitorum*, but he too espoused the same kind of classical approach.

The most astute reader of the *Arithmetica infinitorum*, however, was someone who belonged to a new generation of mathematicians and who immediately recognized its potential: Isaac Newton (1642–1727). He read and made extensive notes on it in the winter of 1664–1665, and his notes continued without interruption as he finished reading and began to explore some of the same themes for himself [Newton, *Papers*, vol. 1, 96–115]. Where Wallis had sought the area of a complete quadrant of a circle, as a numerical ratio to the area of the circumscribed square, Newton set himself the more general and much harder

task of finding partial areas of the quadrant in terms of a free variable, x. The problem was now posed in different language, but Newton tackled it as Wallis had, by interpolating between curves for which the quadrature was easily calculated. Thus he needed, as Wallis had done, to interpolate values in the places marked $*$ in the following sequence:

$$0 * 1 * 3 * 6 * 10 * 15 * \cdots. \tag{12}$$

Wallis had regarded his sequences as being generated by multiplication and so would have written 1, 3, 6, 10 as $1 \times \frac{3}{1} \times \frac{4}{2} \times \frac{5}{3} \times \frac{6}{4} \times \cdots$. Newton saw a much simpler method of interpolating the sequence by means of addition, by noting that it could be written (starting from anywhere in the sequence) as:

$$a \quad a+b \quad a+2b+c \quad a+3b+3c \quad a+4b+6c \quad a+4b+10c \quad \cdots \tag{13}$$

for suitable values of a, b, c. Newton, like Wallis, implicitly assumed continuity and so was able quickly and easily to interpolate the table, and to confirm Wallis's values. Further, because his pattern was simple it was not difficult for him to generalize it to two or more intermediate values, that is to the general coefficients of x^r in the expansion of $(1-x)^{p/q}$. Thus Newton discovered, by a purely interpolative numerical method, the general binomial theorem. Newton's method was different from Wallis's, but it was no coincidence that he wrote his coefficients as Wallis had done, using a sequence of multipliers:

$$\frac{p}{q} \times \frac{p-q}{2q} \times \frac{p-2q}{3q} \times \frac{p-3q}{4q} \times \cdots, \tag{14}$$

or, putting $m = p/q$:

$$m \times \frac{m-1}{2} \times \frac{m-2}{3} \times \frac{m-3}{4} \times \cdots. \tag{15}$$

The possibilities opened up by the discovery of the binomial theorem and associated infinite series expansions can hardly be overestimated. Newton could now write not only any rational function of x as an infinite series, but also trigonometric and logarithmic functions, all of which he could then integrate term by term. He described his results in a handwritten tract entitled 'De analysi per aequationes numero terminorum infinitas', which he sent to Isaac Barrow and John Collins in 1669, and in the 'Epistola prior' and 'Epistola posterior' to Leibniz in 1676 [Newton, *Papers*, vol. 2, 206–247; *Correspondence*, vol. 2, 20–47, 110–163]. When Wallis learned of the contents of the latter he immediately recognized both the extent of Newton's achievement and the debt to the *Arithmetica infinitorum*, and published long extracts from the letters in *A treatise of algebra* in 1685 [Wallis, 1685, 330–346].

With Newton's work the influence and value of the *Arithmetica infinitorum* was no longer in doubt. It would be many years before the ideas of Wallis or Newton would be made fully rigorous, but that did not prevent mathematicians from using the new tools now at their disposal. The book contains the first hints or developments of many of the key concepts of later mathematics: negative and fractional powers, indices, induction, infinitesimals, infinite sums, algebraic formulae for nth terms of sequences, limits, convergence, continuity, the transcendence of π, continued fractions and subscript notation. In many

ways the *Arithmetica infinitorum* was superseded within 20 years of its publication because later developments, particularly of the calculus, rendered Wallis's methods obsolete; but the role of his work in inspiring those developments makes it one of the seminal texts of 17th-century mathematics.

BIBLIOGRAPHY

Andersen, K. 1985. 'Cavalieri's method of indivisibles', *Archive for history of exact sciences, 31*, 291–367.

Brouncker, W. et alii 1658. *Commercium epistolicum de quaestionibus quibusdam mathematicis nuper habitum* (ed. John Wallis), Oxford.

Hobbes, T. 1656. *Six lessons to the professors of mathematiques, one of geometry, the other of astronomy: in the chaires set up by Sir Henry Savile in the University of Oxford*, London.

Huygens, Ch. *Works*. *Oeuvres complètes*, 22 vols., The Hague: Martinus Nijhoff, 1888–1950.

Newton, I. *Correspondence*. *The correspondence of Isaac Newton* (ed. H.W. Turnbull et alii), 7 vols., Cambridge: Cambridge University Press, 1959–1977.

Newton, I. *Papers*. *The mathematical papers of Isaac Newton* (ed. D.T. Whiteside), 8 vols., Cambridge: Cambridge University Press, 1967–1981.

Nunn, T. Percy. 1910–1911. 'The arithmetic of infinites', *Mathematical gazette 5*, 345–357, 378–386.

Panza, M. 2004. 'A l'origine de la notion de nombre transcendant: John Wallis et la quadrature du cercle', in *Newton and the origins of analysis 1664–1666*, Paris: Blanchard, to appear.

Rigaud, S.J. 1841. *Correspondence of scientific men of the seventeenth century*, 2 vols., Oxford: Oxford University Press. [Repr. Hildesheim: Olms, 1965.]

Scott, J.F. 1938. *The mathematical work of John Wallis (1616–1703)*, London: Taylor & Francis. [Repr. New York: Chelsea, 1981.]

Stedall, J.A. 2002. *A discourse concerning algebra: English algebra to 1685*, Oxford: Oxford University Press.

Wallis, J. 1656a. *De sectionibus conicis*, Oxford.

Wallis, J. 1656b. *Due correction for Mr Hobbes, or school discipline, for not saying his lessons right*, Oxford: Oxford University Press.

Wallis, J. 1685. *A treatise of algebra both historical and practical*, London: R. Davis.

Whiteside, D.T. 1961. 'Patterns of mathematical thought in the later seventeenth century', *Archive for history of exact sciences, 1*, 179–388.

CHAPTER 3

CHRISTIAAN HUYGENS, BOOK ON THE PENDULUM CLOCK (1673)

Joella G. Yoder

This is the first modern treatise in which a physical problem is idealized by a set of parameters then analyzed mathematically. It is one of the seminal works of applied mathematics.

First publication. *Horologium oscillatorium, sive de motu pendulorum ad horologia aptato demonstrationes geometricae*, Paris: F. Muguet, 1673. [14] + 161 + [1] pages.

Photoreprints. London: Dawson, 1966. Brussels: Culture et Civilisation, 1966, 1973. Also available on the Gallica website of the *Bibliothèque Nationale*: http://gallica.bnf.fr.

Later edition by W.J. 's Gravesande. In *Christiani Hugenii Zulichemii Opera varia*, 4 vols. in 1, Leiden: J. vander Aa, 1724, 15–192. [Repr. as *Christiani Hugenii Zulichemii opera mechanica, geometrica, astronomica et miscellenea*, 4 vols. in 2, Leiden: G. Potvliet et alia, 1751. The editor used the corrections printed at the end of the 1st ed., replaced many erroneous figures and inserted passages added by Huygens to his personal copy.]

Standard edition. As *Oeuvres complètes*, vol. 18, The Hague: Martinus Nijhoff, 1934, 68–368. [Used 's Gravesande's figures and the corrigenda but returned Huygens's marginalia to the footnotes. Also a facing-page French trans.]

English translation. Christiaan Huygens' the pendulum clock, or, Geometrical demonstrations concerning the motion of pendula as applied to clocks (trans. R.J. Blackwell), Ames: Iowa State University Press, 1986.

German translation. Die Pendeluhr (trans. A. Heckscher and A. von Oettingen), Leipzig: Engelmann, 1913 (*Ostwalds Klassiker der exakten Wissenschaften*, no. 192).

Italian translation. L'orologio a pendolo (trans. C. Pighetti), Florence: Barbèra, 1963. [Also includes *Trattato sulla luce*.]

French translation. L'Horloge oscillante (trans. J. Peyroux), Bordeaux: Bergeret, 1980. [Photorepr. Paris: Blanchard, 1980.]

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

Manuscripts. Main manuscript is missing, but Huygens bequeathed background manuscripts and correspondence to the Library of the University of Leiden, now in the *Codices Hugeniorum*. Much background material is in *Oeuvres complètes*, vols. 17 and 18.

Related article: Newton (§5).

1 THE THREE STRANDS OF HUYGENS'S RESEARCH

The *Horologium oscillatorium* is a superb tapestry woven from the three strands of the science of Christiaan Huygens (1629–1695): mathematics, mechanics, and technology. As is usually the case with scientists who have such a leaning, the young Huygens assimilated his mathematical lessons easily and showed a precocious interest in building small models. He was educated at home by his father, the Dutch poet and diplomat, Constantijn Huygens, and by tutors, including the mathematician Jan Jansz Stampioen de Jonge. He did take courses from Frans van Schooten at the University of Leiden and was sent to the University of Breda to obtain a law degree. Despite his training in mathematics and his natural propensity for science, Christiaan and his brothers were groomed to succeed their father and grandfather as secretaries to the House of Orange that ruled the Netherlands. However, a political crisis in 1650 left the House of Orange deprived of power and Christiaan without a suitable position. Thus, he was free to pursue his interest in science, supported at home by his father. In 1665, Louis XIV invited him to head the newly formed *Académie Royale des Sciences* in Paris, where he resided from 1666 to 1681. He returned home during periods of illness and after the last return, owing to a changed political climate, was never granted permission to rejoin the court in Paris [Bos et alii, 1980].

Mathematics dominated Huygens's early years. During the 1650s he absorbed and extended the results of the Greeks, especially Archimedes. He composed a treatise on floating bodies (published posthumously), related areas under curves to their centers of gravity (*Theoremata de quadratura*, 1651), and devised a better method for approximating pi by inscribed and circumscribed polygons (*De circuli magnitudine inventa*, 1654). Along with his Leiden classmates, Johan Hudde and Johan de Witt, he contributed to the second edition (1659) of van Schooten's Latin translation and gloss on René Descartes's *Géométrie* (§1). During his years in Paris, Huygens turned away from pure mathematics, although he did present to the *Académie* explanations of Pierre de Fermat's methods for finding maxima and minima and for finding tangents. In his final decade, isolated in the Netherlands, he returned to mathematics in response to the correspondence of Gottfried Wilhelm Leibniz.

Huygens's early work in mechanics centered on his confrontations with the theory of Descartes. As he readily admitted, like many young people of his generation he was seduced by the simplicity of Cartesian theory but then discovered yawning holes in the particulars. Nonetheless, throughout his life he adhered to the basic Cartesian view that all physical systems can be reduced to matter in motion, relative motion to be precise. His first major treatise dealt with his analysis of percussion, in which he replaced Descartes's erroneous theorems on colliding bodies with his own corrections based on a strict application of symmetry inherent in the concept of relative motion. Although he finished *De motu corporum ex percussione* by 1656 and presented summaries to both the *Académie*

and the Royal Society of London in the 1660s, Huygens never published the complete work, leaving it for posthumous publication in 1703.

Another area of research inspired by Descartes was dioptrics, or the study of the convergence of light by lenses. Again Huygens differed with his predecessor, particularly over the speed of light and over the refraction of light. Again Huygens moved beyond Cartesian explanations by means of a detailed mathematical analysis. And again, while the topic occupied him throughout his life, yet he published no finished treatise on the subject, though he left a nearly complete manuscript for posthumous publication. Concurrent with his mathematical study of dioptrics, he and his older brother, Constantijn Jr., began grinding lenses and constructing telescopes. With one of their telescopes he discovered Saturn's largest moon, Titan (*De Saturni Luna observatio nova*, 1656), and observed the planet's changing profile, from which he concluded that it was surrounded by a ring (*Systema Saturnium*, 1659). The brothers wrote a description of their lens grinding machine, exchanging the manuscript over successive drafts, but never published it; a Latin translation appeared in the 1703 posthumous edition of Christiaan's works. In Paris, Huygens was outshone by the observational skills of Giovanni Domenico Cassini, but he still advanced certain astronomical topics such as the development of the micrometer. He also became involved in the improvement of the microscope, initially through his father's interest in the work of Antoni van Leeuwenhoek.

2 PENDULUM CLOCKS

Clocks were another topic of research that occupied Huygens throughout his life. He invented his first clock regulated by a pendulum in 1656, and was still working on elaborate variations capable of going to sea in the late 1680s. His marine clocks functioned just well enough to keep him doggedly following that avenue as the solution to the problem of longitude. Ironically, although in one of his clocks he used a spring in place of the driving weight, he never tried using a spring-regulated watch, which he also had invented, at sea because he felt that the influence of temperature on the spring undermined its accuracy. Ultimately, as later inventors would show, the solution was to overcome the temperature variation of a spring and not to try to control the erratic swing of a pendulum.

In 1658 Huygens published a short treatise called simply '*Horologium*'. It described his most recent design of a clock whose timing was regulated by a freely swinging pendulum. Galileo had already conceived of mounting a pendulum to a clock, and his son had even attempted to build one according to a design dictated by his blind father. There are still partisan debates about whether Galileo's clock should be acknowledged as the first successful pendulum clock. As with most mechanical devices, the priority debate devolves into a contest of terms and standards: is the concept enough to ensure precedence or must there be a model that runs reliably and accurately? Huygens argued for priority based on the latter criteria, even claiming that his clock could be used to determine with exactness the inequality of the solar day. Certainly, his book popularized the pendulum clock, and most models were constructed along the lines described in his patent.

A year later, however, Huygens had moved beyond the original design and was planning a second edition of his book that would not only describe the new pendulum clock

but also detail the mathematical and physical underpinnings of its construction. The book, titled *Horologium oscillatorium, sive de motu pendulorum ad horologia aptato demonstrationes geometricae* ('The Pendulum clock, or geometrical demonstrations concerning the motion of pendulums fitted to clocks'), was not published until 1673, and even then it was unfinished. Its appearance at that moment had more to do with politics than science, for Huygens was under pressure to demonstrate his allegiance to Louis XIV during a time when his royal patron was at war with the Netherlands. The dedication to the king is ripe with the expected obsequious compliments, including praise for Louis's generosity even in times of war. To his brother, Huygens confided that he would rush past the bonfires celebrating French victories with downcast eyes.

3 THE FIVE PARTS OF *HOROLOGIUM OSCILLATORIUM*

The contents of Huygens's book are summarized in Table 1. He divided it into five parts that were tightly related. The first described the new clock and how to build it (Figure 1). Many features of the earlier design were carried over, such as the use of a small weight to adjust the timing of the swing, the verge and crown-wheel escapement mechanism for connecting the pendulum to the gears, and the endless cord or chain for winding the clock without stopping it. The radical innovation was that the pendulum no longer swung freely but was limited in its motion by thin metal plates mounted on either side of the pivot point. The plates were bent to a cycloidal shape in the vertical plane, and this part of the book included a description of how to create a cycloid, by rolling a circular object along a straight edge.

At the end of this part, Huygens described a variant of the clock for use at sea, in which the regulating apparatus was essentially stretched out in the horizontal, noncycloidal plane in order to compensate for the tilting of the clock fore and aft. The pendulum was a triangular contrivance that consisted of a cord fastened at either end to a pair of cycloidal plates with the bob hung at the cord's midpoint. Along with the two cycloidal clocks, Part 1 also included a table by which their time could be adjusted for the inequality of the solar day. As Huygens explained, owing to the eccentricity of the earth's orbit and the obliquity of the ecliptic, the length of a day varies and, thus the time given by a clock, which beats at a constant rate and so measures a mean solar day, must be adjusted if it is to be scientifically accurate. This adjustment for the Equation of Time, as it is called, was essential if the clocks were to be used for astronomical observation, on land or at sea.

The primary purpose of the next two parts of the book was to prove that the cycloidal pendulum was isochronous; that is, the bob would complete its swing at a uniform rate independent of the magnitude of the swing. Thus, the clock would keep exact time irrespective of anomalies in the pendulum's oscillation. Despite the enduring myth, Galileo's claim that a freely swinging pendulum is isochronous is wrong, although if the pendulum is made to swing through very small arcs, as it does in a grandfather clock or in Huygens's 1658 clock, the variation is negligible.

Part 2 comprised a set of propositions on fall, beginning with the hypothesis of rectilinear inertial motion and then introducing compound motion owing to gravity. The object in motion was an idealized point mass; gravity was assumed and not explained. In this section Huygens took up Galileo's analysis of free fall and fall along inclined planes and extended

Table 1. Contents by grouped propositions of Huygens's book. 161 pages.

Pt., Props.	**Topics**
I	Description of a clock regulated by a cycloidal pendulum; table for equation of time; marine clock.
II, 1–8	Bodies falling freely and through inclined planes [Galilean section].
II, 9–11	Fall, and subsequent ascent, in general.
II, 12–15	Tangent of cycloid; history of the tangent problem; generalization to similar curves.
II, 16–26	Fall through cycloid.
III, 1–4	Definition of evolute and its companion; their relationship.
III, 5–6, 8	Evolute of cycloid and parabola.
III, 7, 9a	Rectification of cycloid and semicubical parabola; history of the problem.
III, 9b–e	Circles equal in area to surfaces of conoids; rectification of parabola equivalent to quadrature of hyperbola; approximation of the latter by logarithms.
III, 10	Evolutes of ellipses and hyperbolas; rectifications of those evolutes.
III, 11	Evolute of any given curve; rectification of that evolute; examples.
IV, 1–6	Simple pendulum equivalent to a pendulum compounded of weights along its length.
IV, 7-20	Center of oscillation of a plane figure and its relationship to center of gravity.
IV, 21–22	Centers of oscillation of common plane and solid figures.
IV, 23–24	How to adjust a pendulum clock using a small weight; application to a cycloidal pendulum.
IV, 25	Universal measure of length based on seconds pendulum.
IV, 26	Constant of gravitational acceleration.
V	Description of a clock regulated by a rotating pendulum; 13 theorems on centrifugal force stated without proofs.

it to fall along curves, where the curve was approximated at each point by its tangent plane. Although completely general, the propositions were background for the main theorem in which Huygens demonstrated that, given a cycloid erected in the vertical plane with its axis perpendicular and its vertex at the bottom, no matter where along the inverted cycloid a body is released it will reach the bottom in a fixed amount of time; that is, fall along an inverted cycloid is isochronous. In modern terms, Huygens showed that the time of fall is a constant $(\pi/2)\sqrt{(2D/g)}$ seconds, where D is the diameter of the circle that generated the cycloid (and $2D$ is the length of the pendulum in Part 1). Or, to phrase the conclusion using proportions, as Huygens did, the time of fall from any point along the cycloid is to

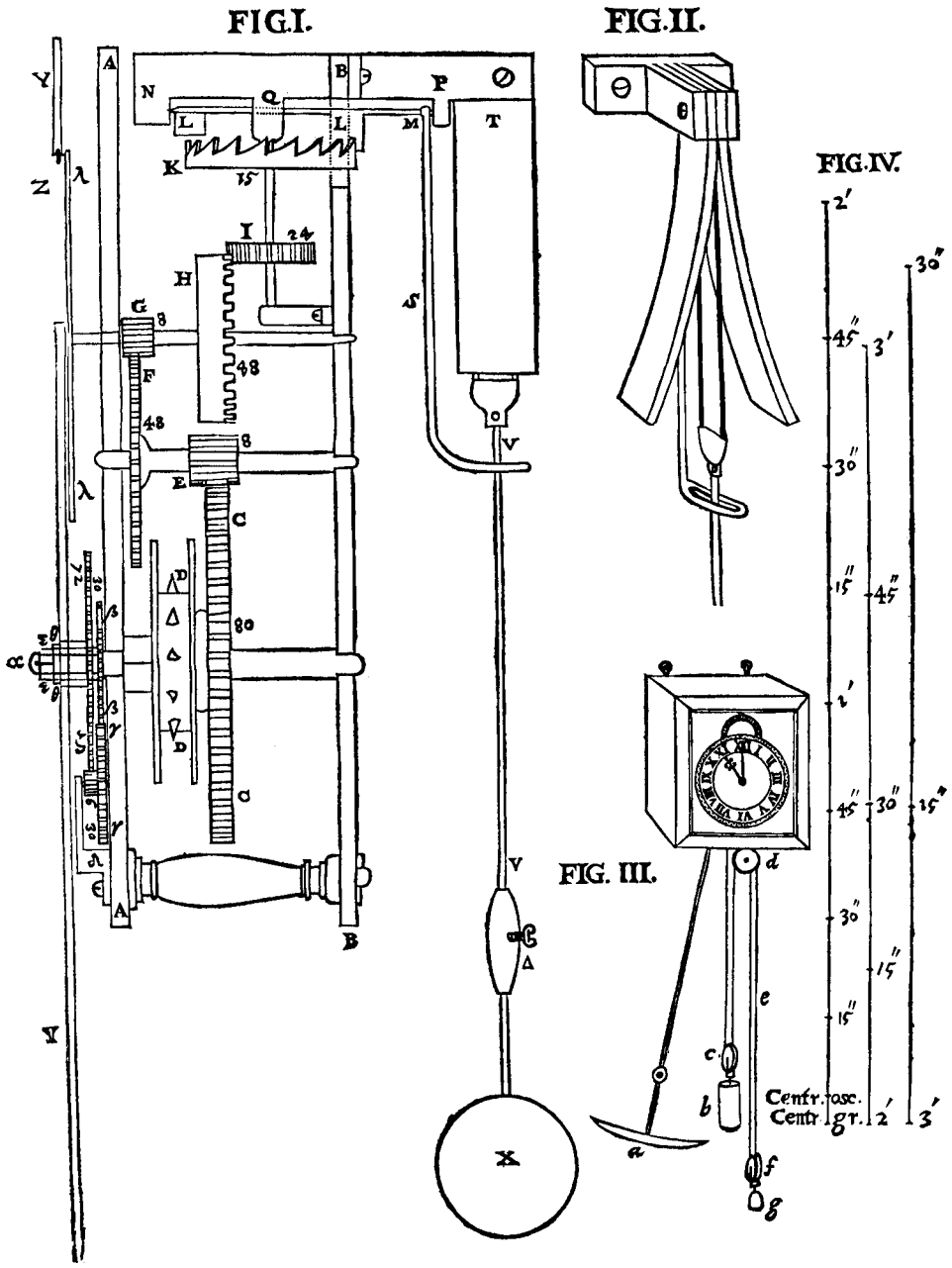


Figure 1. Huygens's diagrams for his clock design.

the time of free fall through the diameter of the generating circle as a semicircumference of a circle is to its diameter.

In Part 3 Huygens went further into the mathematics embodied in his clock. He introduced the concept of an evolute, a curve that is 'unrolled' (*evolutus* in Latin) to create a second curve, which Huygens called 'that described by the unrolling' and later mathematicians labeled either 'evolvent' or 'involute'. Although in this section he never discussed the technological application behind his theory, the pendulum of his clock served as his model: if a cord were 'unrolled' off of one curved plate, the bob would trace out the involute. Given the results of Part 2, he obviously wanted the curve swept out by the bob to be a cycloid. When he first discovered the isochronism of the cycloid in 1659, he determined its evolute using infinitesimal techniques. He took two points very close together on the cycloid, derived the point of intersection of the perpendiculars through them, and deduced the curve defined by all such points. That evolute would be the shape into which the curved plates should be bent. Imagine his surprise to discover that it was another cycloid. For the published derivation, Huygens banished infinitesimals and proved that the evolute of a cycloid is another cycloid using a sequence of propositions that defined the reciprocal relationship between the tangent of the evolute and the perpendicular (we would now say normal) of the involute.

In Part 4 Huygens finally introduced physical parameters into his analysis by addressing the problem of the compound pendulum [Gabbey, 1982]. He began by explicitly stating the definitions of a pendulum, as any figure that can continue reciprocal motion around a point or axis by means of its own weight; of a simple pendulum, as one that has a weightless cord and a point mass bob; and of a compound pendulum, as any suspended object with weight distributed throughout. Further definitions included what he meant for two pendulums to be isochronous (they were to swing through equal arcs in equal times), from which he could then speak of the center of oscillation of a suspended object as the point along the axis at which to situate the bob of a simple pendulum that would be isochronous with the compound one. A few more definitions were followed by the fundamental hypothesis of this passage, usually referred to as Torricelli's Principle: a system of weights cannot rise of its own accord above its center of gravity. Applying that hypothesis, Huygens was able to derive a simple pendulum that was isochronous to a given compound pendulum, first for one consisting of a set of weights distributed along a pendulum's cord, then for a general figure, and then, as a concluding tour de force, for a whole set of plane and solid figures. Commentators have noted that Huygens did not break free of gravity and deal with a generalized system of masses, most particularly a rotating body.

At last, as the fourth part ended, Huygens drew together all of his evidence and applied it to the clock described in Part 1. Parts 2 and 3 had justified the introduction of the cycloidal plates to his clock by mathematically demonstrating that the bob of a pendulum that banks along cycloidal plates as it swings will trace out a cycloid. Thus the clock described in Part 1 was theoretically isochronous. Part 4 gave him a way to account for the mass of the cord and bob and to adjust the rate of swing of the pendulum by means of a small weight that could be moved up and down the cord. Thus the clock could be made physically as exact as possible. Frustrated with international variations in the length of a foot, Huygens next proposed a universal measure based on this very accurate clock. He defined a clock-foot to be equal to one-third the length of a simple pendulum that was isochronous with

a cycloidal-pendulum clock that had been previously adjusted to beat precisely once a second. Ever the perfectionist, Huygens exposed the flaw in his beautiful cycloidal clock: all his theorems had dealt with a pendulum of fixed length but, as the cycloidal pendulum banked along its cycloidal plates, its cord was foreshortened. However, he rationalized, it did not introduce that large an error.

In the last proof of the book, Huygens finally returned to the problem that had instigated his work on the cycloidal pendulum: find the distance traversed in a given time by a body falling perpendicularly under the influence of gravity. If the time is one second, the distance fallen from rest is numerically one-half the constant of gravitational acceleration (the modern formula is $d = 1/2gt^2$). By the last proposition of Part 2, the time of free fall can be related to the length (3 clock-feet) of a cycloidal pendulum that swings one arc a second. Inverting the concluding proportion of that proposition yields: the time a body falls freely through one-half the length of the pendulum (18 inches = the diameter of the generating circle of the cycloid) is to one-half a second (the time it takes the pendulum to reach bottom = half its swing) as the diameter (d) of a circle is to its semicircumference ($\pi d/2$). Huygens approximated π by $355/113$; therefore the time of fall through 18 inches is $(1/2) \times 2 \times (113/355)$ seconds or 19.1 thirds (a sixtieth of a second). The Galilean formula for relating distance fallen to time-squared gives:

$$\text{distance fallen in one second (60 thirds) is to 18 inches as } (60)^2 \text{ is to } (19.1)^2, \quad (1)$$

yielding a distance of 14 feet and 9.6 inches.

At last the modern reader can deduce the value of the constant of gravitational acceleration, which Huygens never expressed directly. It is twice the distance Huygens did derive, or 29 feet, 7.2 inches in clock-feet, which is about 987 cm per sec^2 in the units that eventually did become a universal measure. Yielding to contemporary measure for his Parisian locale, Huygens converted his value to 15 Parisian feet, one inch.

In an odd anticlimax, Huygens then proceeded to describe an experiment to find the constant using a pendulum not attached to a clock but rather fixed to a wall and with its bob attached to a lead weight by a thin cord, which then was to be severed by a flame, leaving the two objects to swing/fall freely simultaneously. The bob of the pendulum, which was to be blackened with soot, was to hit a paper scale dragged along the wall by the falling weight. Strange as it sounds, Huygens probably executed just such an experiment, because he performed one similar to it in 1659 and determined a value very close to that derived from the clock.

In Part 5 of *Horologium oscillatorium* Huygens introduced another clock, this one based on a three-dimensional mathematical model that paralleled the design elements in the clock of Part 1. Of course, the earlier clock was three-dimensional, but Huygens had treated it as a two-dimensional system; the pendulum was presumed to swing in a plane and the curved plates along which it banked were shaped only for that plane (in the other dimension they were straight). The progenitor of this second clock had also begun life with a freely swinging pendulum, but its bob had rotated in a circle and hence the cord traced out a cone. But as with the cycloidal clock, he modified the clock of Part 5 so that it was constrained by a curved plate to move isochronously. In one of his discoveries made in 1659, Huygens had shown that a ball circulating inside a chalice shaped like a paraboloid (the surface generated

by revolving a parabola about its axis) completed any circle in the same amount of time, regardless of how high or low the ball was in the chalice; the paraboloid was isochronous. With his theory of evolutes, he could easily transfer his result to the rotating pendulum and deduce the proper shape for a curved plate that would hold the revolving bob onto the imaginary surface of a paraboloid. By Part 3, Proposition 8, the evolute of a parabola is a semicubical parabola. So, if a pendulum were mounted to a curved plate shaped like a semicubical parabola and if the combined pendulum and plate were then rotated about an axis, the bob would isochronously sweep out a paraboloid. Like its cycloidal brother, the paraboloidal clock would theoretically keep perfect time.

With the description of the paraboloidal pendulum clock, *Horologium oscillatorium* ended. Huygens did not continue on and provide a detailed mathematical study to accompany the second clock which would parallel the analysis for the first. Instead, he merely appended a list of thirteen theorems regarding motion in a circle that had guided his creation of the clock in 1659, but he withheld their proofs. Huygens's stated intent was to save the results for a larger work that would present his definitive explanation of circular motion and centrifugal force. Alas, that work never materialized, although hints of it are found in his manuscripts. The original proofs finally appeared in the 1703 posthumous edition under the title *De vi centrifuga*. Even without the added material, *Horologium oscillatorium* is a masterpiece of mathematical physics, but in holding back his work on circular motion Huygens further undermined his standing and influence.

Many of the propositions in the book had nothing to do with the clock but everything to do with the history of Huygens's ideas. He had begun his mathematical study of pendulums when an attempt to measure the gravitational constant using a pendulum failed to give consistent results. Rather than continue the experiment, he idealized it into a mathematical study comparing free fall and fall along a circle. His first attempt to solve the mathematical problem by essentially retracing Galileo's work on fall ended in failure at the point where he tried to extend his results to curvilinear fall. He abandoned the Galilean approach and tackled the problem directly by using a very personal form of infinitesimal analysis, a strange fusion of the Cartesian analytic geometry that he had learned from van Schooten with classical geometry and contemporary infinitesimal techniques. Significantly, when he came to publish the results, he reverted to a strictly classical presentation that built upon his Galilean study.

4 ON HUYGENS'S MATHEMATICAL STYLE

Parts 2 and 3 summarized many of the achievements of 17th-century mathematics, particularly Huygens's own contributions. They were written in the classical style of Archimedean geometry, with every step strictly substantiated by an appropriate proposition. They were written in the classical language of proportions, with nothing displayed in easily recognizable formulas. In a quintessential tribute to Archimedes, Huygens paused in the middle of Part 3 to reduce the areas of surfaces of conoids to the areas of circles. The results paralleled the propositions given in his early *Theoremata de quadratura* of 1651, in which he had related areas of conics to their centers of gravity.

Despite his classicism, Huygens was at pains to place his discoveries in the context of contemporary priority debates that accompanied the advances in quadrature (the determi-

nation of the area of a curved figure) and rectification (the determination of the length of a curve), all material that set the stage for the growth of the calculus. In Part 2 he presented his method for finding the tangent to the cycloid and reviewed the history of earlier methods. In Part 3 he asserted his claim that he was at the forefront of discoveries regarding the rectification of important curves, including his beloved cycloid. In fact, he was able to develop a general method of rectification based on evolutes. After all, as the cord unwound off the evolute, it literally straightened, or rectified, the curve. But the method had a major flaw, for it rectified the evolute of a given curve, not the curve itself. He closed Part 3 with derivations of the evolutes of the ellipse, hyperbola, and higher order conics accompanied by the rectifications of those evolutes. More significant was his reduction of the rectification of the parabola to the quadrature of the hyperbola, which he then solved by numerical approximation using logarithms. In Part 4, ever seeking his due, he sketched his abbreviated history of the problem of the center of oscillation, particularly referring to his early exposure to the subject via his youthful correspondence with Marin Mersenne.

5 UNFOCUSED RECEPTION

Because much of the information in the *Horologium oscillatorium* had become known during its 14-year gestation, its reception was piecemeal. While clerking for diplomatic missions in 1660, Huygens had already shown off a clock based on the cycloidal model in both Paris and London. Indeed, his reputation as a designer of clocks is one reason that he later received appointment to the *Académie*. Very few clocks were actually built to the cycloidal design, and Huygens even admitted that a free pendulum with a small swing would suffice. But he was adamant about the superiority of his design as a scientific instrument. Alas, before he died he learned that his carefully calibrated clocks did not beat the same in every location when members of the *Académie* took them along on distant expeditions. So confident was he that the clocks were accurate that he correctly surmised that the shape of the Earth must be affecting the timing. The constant of gravitational acceleration was not, it seems, quite constant.

The reception of the mechanical theory was also diffused by the delay in publication. The isochronism of the cycloid became known along with the clock that embodied it. Lord William Brouncker even made several attempts to prove the result before Huygens published his own. Huygens's derivation of the center of oscillation for compound pendulums engendered a debate in the *Académie* when he presented a draft version to his colleagues. The major objection concerned his fundamental hypothesis, the Torricelli Principle. Huygens revised sections of Part 4 in response to criticisms, particularly those of Gilles Personne de Roberval. Still, objections continued after publication. Eventually Jakob Bernoulli replaced the fundamental hypothesis with the law of the lever, proved that the center of oscillation is equal to the center of percussion, and extended the results to rotating bodies (1703). Moreover, Huygens's studies of collisions and of compound pendulums became elements of the *vis viva* debate at the turn of the century, as partisans tried to define precisely what they meant by force, motion, and conservation of motion.

Huygens's early work on circular motion became absorbed into the larger debate over the vortex theory and the cause of gravity. Before going to Paris, he had tended to treat

mechanics as a strictly mathematical system. But once in Paris and confronted with opponents of Cartesianism, he was forced to assess the foundations of his mechanical beliefs, most particularly with respect to the vortex explanation of centrifugal force. Presumably Huygens withheld publishing the proofs of his theorems on circular motion in *Horologium oscillatorium* because he wanted to include them in a broader study of motion that would codify his position on vortices. The task ballooned when Huygens realized he had to respond to Newton's *Principia,* which essentially demolished Cartesian vortices. Ironically, Newton claimed to be modeling his masterpiece on *Horologium oscillatorium* and he saluted its author by calling his new force 'centripetal' (compare §5.2.1).

The influence of Huygens's mathematical theory of evolutes was fleeting but important. First, he himself applied evolutes to optics in order to derive the wave front of light as it moved through Iceland spar. Indeed, he thought that his explanation of the strange double refraction that occurs was strong proof that light must move in a wave. In particular, he gave the definition of a wave front that is now referred to as Huygens Principle; namely, as the common tangent (envelope) to all the secondary waves that emanate from the previous position of the wave. Moreover, he extended this approach by finding the wave fronts (called caustics) formed by reflection and refraction in a spherical surface, where light rays do not converge. The Bernoulli brothers generalized this work to other nonconvergent cases.

Huygens presented his theory in *Traité de la lumière* (1691). Published jointly with that work was his *Discours de la cause de la pesanteur,* which summarized his latest thoughts on circular motion. Even with that, however, he did not complete the work promised by *Horologium oscillatorium*, Part 5. In addition, in Part 1, Huygens had remarked that he did not know if any other curve besides the cycloid was its own evolute. By 1678 he himself discovered that the epicycloid also replicated itself when 'evolved', and in 1692 Jakob Bernoulli showed that the logarithmic spiral likewise qualified.

Implicit in the theory of evolutes is the mathematical concept of the radius of curvature. At any point an involute can be approximated by a circle centered on its evolute, where the 'unrolled' line connecting the two curves is the radius of that circle and hence, by definition, is the radius of curvature.

It is debatable whether Huygens really thought of a curve as being continuously measured in that way, especially since he was just as likely to approximate a curve by a parabola as by a circle. However, when Leibniz more formally introduced the 'osculating' circle, Huygens's reaction was to claim priority because his general derivation of an evolute was similar. Indeed, it was, as his proof of Part 2, Proposition 11 showed. In this general case, Huygens reverted to the infinitesimal approach that he had used originally to derive the evolute of the cycloid. From the relationship of involute and evolute, any two normals to the curve *ABF*, such as *BMD* and *FNE*, are tangent to the evolute at D and E (Figure 2). If B and F are indefinitely close, then both D and E can be approximated by the point of intersection, G, of the two normals and the curve between B and F can be treated as a line segment, namely as the extension of the tangent, *BH*, to *ABF* at B. Draw *FPL* and *BK* perpendicular to the axis *AL*, and draw *BPO* perpendicular to *FPL*. By the similarity of triangles *BOG* and *MNG*, $BG/MG = BO/MN$. Huygens then decomposed this proportion into $BG/MG = BO/BP \times BP/MN$. Since $BP = KL$ and, by similar triangles,

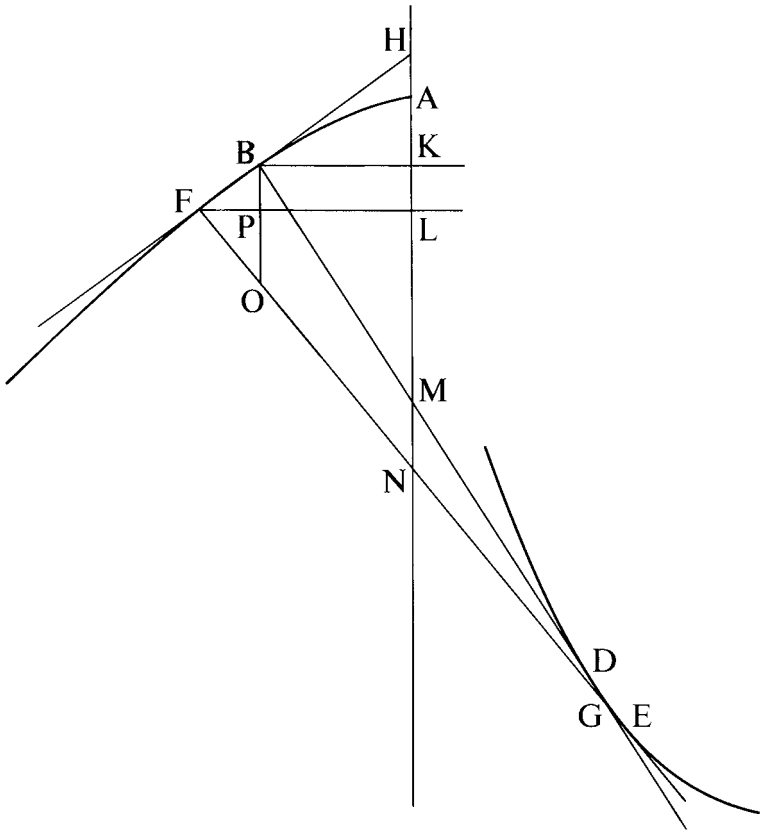


Figure 2. Huygens's derivation of an evolute.

$BO/BP = HN/HL$, the proportion can be rewritten as

$$BG/MG = HN/HL \times KL/MN. \tag{2}$$

This composed proportion was Huygens's fundamental formula for deriving any point, G, on the evolute to the curve ABF. It is reducible to the modern formula for the radius of curvature for a twice-differentiable curve. However, only with Jakob Bernoulli do we get an analytical expression that represents the rate of change of the primary curve $((ds/dx)^3/(d^2y/dx^2))$. Did Huygens determine curvature? As with priority claims about the pendulum clock, the answer depends upon the criteria.

The lasting importance of *Horologium oscillatorium* stemmed more from its applied mathematics than from its pure mathematics. The next generation of mathematicians spent a great deal of time trying to find curves that satisfied specific physical properties. What other curve, if any, is a tautochrone (curve in which fall is isochronous)? What curve does a hanging chain delineate? What shape does a sail take? What is the curve of fastest descent? These were the test cases for the new mathematical technique Leibniz called 'calculus'.

While he was alive, Huygens was still able to find many of the solutions using his own *ad hoc* infinitesimal geometrical methods. But the younger generation reduced it all to formulas and then generalized those.

Foremost, Huygens gave us precise time. His clocks were the first timekeepers to be accurate enough to be reliable in scientific experiments. Once time could be accurately measured, other variables could be graphed against time. From that foundation, one could then proceed to consider instantaneous variation. Although he himself did not go down that path, Huygens opened the way with his exacting mathematical dissections of physical problems into a minimum of parameters. Others, especially the Bernoullis, built on his examples. In *Horologium oscillatorium* Huygens extolled the cycloid, citing 'the power of this line to measure time'. What a remarkable statement of the mathematical reductionism this made him such an important figure in the foundation of applied mathematics.

BIBLIOGRAPHY

Bell, A.E. 1947. *Christian Huygens and the development of science in the seventeenth century*, London: Arnold.

Bos, H.J.M. 1974. 'Huygens, Christiaan', in *Dictionary of scientific biography*, vol. 6, 597–613.

Bos, H.J.M., Rudwick, M.J.S., Snelders, H.A.M., and Visser, R.P.W. (eds.) 1980. *Studies on Christiaan Huygens: invited papers from the symposium on the life and work of Christiaan Huygens, Amsterdam, 22–25 August 1979*, Lisse: Swets & Zeitlinger.

Edwardes, E.L. 1977. *The story of the pendulum clock*, Altrincham: Sherratt. [Includes English trans. of the 1658 *Horologium*.]

Elzinga, A. 1972. *On a research program in early modern physics*, Göteborg: Akademiförlaget.

Gabbey, A. 1982. 'Huygens et Roberval', in *Huygens et la France: Table ronde du Centre National de la Recherche Scientifique, Paris, 27–29 mars 1979*, Paris: Vrin, 69–84. [Includes primary material regarding Part 4. See also other articles there.]

Huygens, Ch. *Oeuvres complètes de Christiaan Huygens, publiées par la Société hollandaise des sciences*, 22 vols., The Hague: Martinus Nijhoff, 1888–1950.

Mormino, G. 1993. *Penetralia motus: La fondazione relativistica della meccanica in Christiaan Huygens, con l'edizione del 'Codex Hugeniorum 7A'*, Florence: La Nuova Italia.

Palm, L. (ed.) 1996. *Christiaan Huygens*, Hilversum: Verloren. [A thematic issue of *De zeventiende eeuw* devoted to papers presented at a conference on Huygens in Leiden in 1995.]

Vilain, Ch. 1996. *La mécanique de Christian Huygens: La relativité du mouvement au XVII^{e} siècle*, Paris: Blanchard.

Yoder, J.G. 1988. *Unrolling time: Christiaan Huygens and the mathematization of nature*, Cambridge: Cambridge University Press.

CHAPTER 4

GOTTFRIED WILHELM LEIBNIZ, FIRST THREE PAPERS ON THE CALCULUS (1684, 1686, 1693)

C.S. Roero

The invention of the differential and integral calculus, in these papers printed in the *Acta Eruditorum*, is one of the most important and revolutionary developments in mathematics. Leibniz and Newton share out the glory of the invention of the infinitesimal calculus, that they found independently. The priority of publication is due to Leibniz, who had the fortune to be followed by mathematicians of first rank who collaborated on the diffusion of his methods.

First publications.

a) 'Nova methodus pro maximis et minimis, itemque tangentibus, quae nec fractas, nec irrationales quantitates moratur, et singulare pro illis calculi genus', *Acta Eruditorum*, (1684), 467–473 + Tab. xii. October issue.

b) 'De geometria recondita et analysi indivisibilium atque infinitorum', *Acta Eruditorum*, (1686), 292–300. June issue.

c) 'Supplementum geometriae dimensoriae, seu generalissima omnium tetragonismorum effectio per motum: similiterque multiplex constructio lineae ex data tangentium conditione', *Acta Eruditorum*, (1693), 385–392. September issue.

Photoreprint of a). In P. Dupont and C.S. Roero, *Leibniz 84. Il decollo enigmatico del calcolo differenziale*, Rende (CS): Mediterranean Press, 1991, 154–161.

Reprints of a). In *Historia fluxionum, sive tractatus originem* [. . .] *exhibens* (ed. J. Raphson), London: Pearson, 1715, 19–26. Also in [G. G. L.], *Nova methodus* [. . .], *Opuscula omnia Actis Eruditorum* [. . .], Venice: J.B. Pasquali, vol. 1, 1740, 270–275.

Reprints of a)–b). In *Die Werke von Jakob Bernoulli*, vol. 5 (ed. A. Weil and M. Mattmüller), Basel: Birkhäuser, 1999, 13–27.

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

Reprints of a)–c). In Leibniz, *Opera omnia* (ed. L. Dutens), vol. 3, *Opera mathematica*, Geneva: de Tournes, 1768, 167–172, 188–194, 287–293. Also in *Leibnizens mathematische Schriften* (ed. C.I. Gerhardt), vol. 5, Halle: Schmidt, 1858 (photorepr. Hildesheim: Olms, 1971), 220–226, 226–233, 294–301.

Partial English translation of a)–c) by D.J. Struik in his *A source book in mathematics 1200–1800*, Cambridge, MA: Harvard University Press, 1969, 272–280, 281–282 (part), 282–284 (part). [a) repr. in J. Fauvel and J. Gray (eds.), *The history of mathematics: a reader*, London: Macmillan, 1987, 428–434.]

Partial English translation of a) *and* b) by E. Walker in D.E. Smith (ed.), *A source book in mathematics*, vol. 2, New York: McGraw-Hill, 1929 (repr. New York: Dover, 1959), 620–623, 624–626.

German translation of a) *and* c) by G. Kowalewski in his (ed.), *Leibniz über die Analysis des Unendlichen*, Leipzig: Engelmann, 1908 (*Ostwald's Klassiker der exakten Wissenschaften*, no. 162), 3–11, 72–76; 24–34, 79.

French translation of a) by P. Mansion in *Mathesis, 4* (1884), 177–185. [Repr. in his *Résumé du cours d'analyse infinitésimale de l'Université de Gand*, Paris: Gauthier–Villars, 1887, 199–208.]

French translation of a)–c) by M. Parmentier in his (ed.), *G.W. Leibniz, La naissance du calcul différentiel*, Paris: Vrin, 1989, 96–117, 131–143, 252–267.

Italian translations of a). 1) By E. Carruccio in *Periodico di matematiche, (4) 7* (1927), 285–301. [Repr. in G. Castelnuovo, *Le origini del calcolo infinitesimale nell'era moderna*, Milano: Feltrinelli, 1938 (repr. 1962), 163–177.] 2) By C.S. Roero in P. Dupont and Roero, *Leibniz 84. Il decollo enigmatico del calcolo differenziale*, Rende (CS): Mediterranean Press, 1991, 23–49.

Partial Italian translation of b) and c) by L. Giacardi and C.S. Roero, in P. Dupont, *Appunti di storia dell'analisi infinitesimale*, vol. 2, part 2, Turin: Cortina, 1982, 862–864, 873–876.

Russian translation of a) by A.P. Jushkevich in *Zhurnal uspechi matematicheskich nauk, 3* (1948), 166–173.

Spanish translations of a) *and* b). 1) By J. Babini in his (ed.), *El calculo infinitesimal*, 1972, 41–51. 2) By T. Martin Santos in Leibniz, *Análisis infinitesimal*, Madrid: Tecnos, 1987, 3–29.

Related articles: Descartes (§1), Newton (§5), Berkeley (§8), Euler on the calculus (§14).

1 LEIBNIZ'S RESEARCH ON THE INFINITESIMAL MATHEMATICS

The invention of the Leibnizian infinitesimal calculus dates from the years between 1672 and 1676, when Gottfried Wilhelm Leibniz (1646–1716) resided in Paris on a diplomatic mission. In February 1667 he received the doctor's degree by the Faculty of Jurisprudence

of the University of Altdorf and from 1668 was in the service of the Court of the chancellor Johann Philipp von Schönborn in Mainz. At that time his mathematical knowledge was very deficient, despite the fact that he had published in 1666 the essay *De arte combinatoria*. It was Christiaan Huygens (1629–1695), the great Dutch mathematician working at the Paris Academy of Sciences, who introduced him to the higher mathematics. He recognised Leibniz's versatile genius when conversing with him on the properties of numbers propounded to him to determine the sum of the infinite series of reciprocal triangular numbers. Leibniz found that the terms can be written as differences and hence the sum to be 2, which agreed with Huygens's finding. This success motivated Leibniz to find the sums of a number of arithmetical series of the same kind, and increased his enthusiasm for mathematics. Under Huygens's influence he studied Blaise Pascal's *Lettres de A. Dettonville*, René Descartes's *Geometria* (§1), Grégoire de Saint-Vincent's *Opus geometricum* and works by James Gregory, René Sluse, Galileo Galilei and John Wallis.

In Leibniz's recollections of the origin of his differential calculus he relates that reflecting on the arithmetical triangle of Pascal he formed his own harmonic triangle in which each number sequence is the sum-series of the series following it and the difference-series of the series that precedes it. These results make him aware that the forming of difference-series and of sum-series are mutually inverse operations. This idea was then transposed into geometry and applied to the study of curves by considering the sequences of ordinates, abscissas, or of other variables, and supposing the differences between the terms of these sequences infinitely small. The sum of the ordinates yields the area of the curve, for which, signifying Bonaventura Cavalieri's 'omnes lineae', he used the sign '$\int$', the initial letter of the word 'summa'. The difference of two successive ordinates, symbolized by 'd', served to find the slope of the tangent. Going back over his creation of the calculus Leibniz wrote to Wallis in 1697: 'The consideration of differences and sums in number sequences had given me my first insight, when I realized that differences correspond to tangents and sums to quadratures' [Gerhardt, 1859, 25].

The Paris mathematical manuscripts of Leibniz published by Gerhardt [1846, 1855, 1863], translated into English in [Child, 1920] and discussed in detail by Hofmann [1949, 1974] show Leibniz working out these ideas to develop an infinitesimal calculus of differences and sums of ordinates by which tangents and areas could be determined and in which the two operations are mutually inverse. The reading of Blaise Pascal's *Traité des sinus du quart de circle* gave birth to the decisive idea of the characteristic triangle, similar to the triangles formed by ordinate, tangent and sub-tangent or ordinate, normal and sub-normal [Gerhardt, 1858, 399–400]. Its importance and versatility in tangent and quadrature problems is underlined by Leibniz in many occasions, as well as the special transformation of quadrature which he called the transmutation theorem by which he deduced simply many old results in the field of geometrical quadratures [Bos, 1980, 62–65]. The solution of the 'inverse-tangent problems', which Descartes himself said he could not master, provided an ever stronger stimulus to Leibniz to look for a new general method with optimal signs and symbols to make calculations simple and automatic.

2 THE ENIGMATIC FIRST PUBLICATION

The first public presentation of differential calculus appeared in October 1684 in the new journal *Acta Eruditorum*, established in Leipzig, in only six and an half pages, written in a disorganised manner with numerous typographical errors. In the title, 'A new method for maxima and minima as well as tangents, which is impeded neither by fractional nor irrational quantities, and a remarkable type of calculus for them', Leibniz underlined the reasons for which his method differed from—and excelled—those of his predecessors. In his correspondence with his contemporaries and in the later manuscript 'Historia et origo calculi differentialis', Leibniz predated the creation of calculus to the Paris period, declaring that other tasks had prevented publication for over nine years following his return to Hannover [Gerhardt, 1846, 4–6, 14–17; 1858, 395–398, 404–407].

Leibniz's friends Otto Mencke and Johann Christoph Pfautz, who had founded the scientific journal *Acta Eruditorum* in 1682 in Leipzig, encouraged him to write the paper; but it was to be deemed very obscure and difficult to comprehend by his contemporaries. There is actually another more urgent reason which forced the author to write in such a hurried, poorly organised fashion. His friend Ehrenfried Walter von Tschirnhaus (1651–1708), country-fellow and companion of studies in Paris in 1675, was publishing articles on current themes and problems using infinitesimal methods which were very close to those that Leibniz had confided to him during their Parisian stay [Tschirnhaus, 1682; 1683a, 122–124; 1683b, 433–437]; Leibniz risked having his own invention stolen from him [Hess, 1986, 73]. The structure of the text, which was much more concise and complex than the primitive Parisian manuscript essays, was complicated by the need to conceal the use of infinitesimals. Leibniz was well aware of the possible objections he would receive from mathematicians linked to classic tradition who would have stated that the infinitely small quantities were not rigorously defined, that there was not yet a theory capable of proving their existence and their operations, and hence they were not quite acceptable in mathematics.

Leibniz's paper opened with the introduction of curves referenced to axis x, variables (abscissas and ordinates) and tangents. The context was therefore geometric, as in the Cartesian tradition, with the explicit representation of the abscissa axis only. The concept of function did not yet appear, nor were dependent variables distinguished from independent ones. The characteristics of the introduced objects were specified only in the course of the presentation: the curve was considered as a polygon with an infinity of infinitesimal sides (that is, as an infinitangular polygon), and the tangent to a point of the curve was the extension of an infinitesimal segment of that infinitangular polygon that represented the curve. Differentials were defined immediately after, in an ambiguous way. Differential dx was introduced as a finite quantity: a segment arbitrarily fixed *a priori*, which is even shown in Figure 1. This definition however would never be used in applications of Leibniz's method, which was to operate with infinitely small dx in order to be valid. The ordinate differential was introduced apparently with a double definition: 'dv indicates the segment which is to dx as v is to XB, that is, dv is the difference of the v'.

In the first part Leibniz establishes the equality of the two ratios ($dv : dx = v : \mathrm{XB}$), the equality deduced by the similitude between the finite triangle formed by the tangent, the ordinate and the subtangent, and the infinitesimal right-angle triangle whose sides are

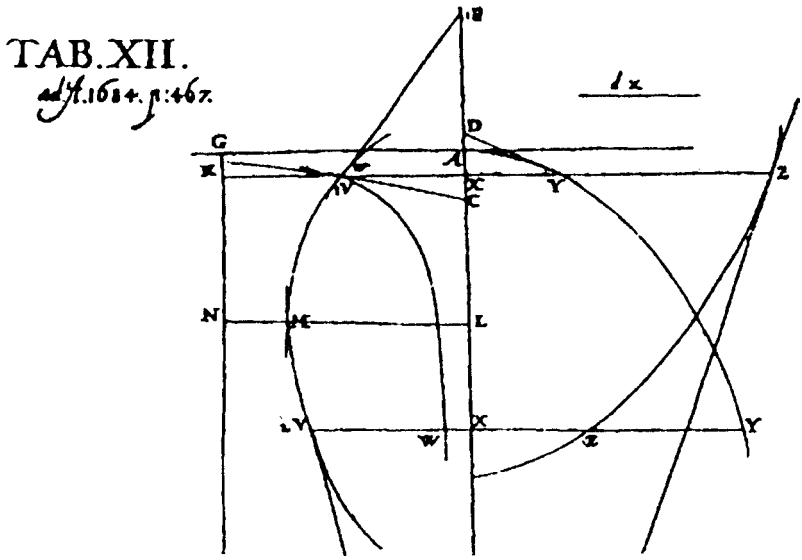


Figure 1. The top half of Leibniz's principal diagram for his first paper.

the differentials thereof and is called 'characteristic triangle'. But the proportion contains a misprint in the expression for the subtangent that would be corrected only in the general index of the first decade of the journal [*Acta Eruditorum*, 1693], 'Corrigenda in Schediasmatibus *Leibnitianis*, quae Actis Eruditorum Lipsiensibus sunt inserta'). The second part ('dv is the difference of the v') mentioned the difference between the two ordinates which must lie infinitely close: $dv = v(x + dx) - v(x)$. In actual fact, the proportion was needed to determine the tangent line and the definition of dv was consequently the second, as explicitly appeared in three of Leibniz's Parisian manuscripts [Gerhardt, 1855, 140–155; trans. Child, 1920, 124–144]. Considering the corresponding sequences of infinitely close abscissas and ordinates, Leibniz called differentials into the game as infinitely small differences of two successive ordinates (dv) and as infinitely small differences of two successive abscissae (dx), and established a comparison with finite quantities reciprocally connected by the curve equation [Dupont and Roero, 1991, 65–71].

These first concepts were followed, without any proof, by differentiation rules of a constant a, of ax, of $y = v$, and of sums, differences, products and quotients. For the latter, Leibniz introduced double signs, whereby complicating the interpretation of the operation. In [*Acta Eruditorum*, 1693] he would decide to abolish 'ambiguous signs' in divisions and their interpretation. The proofs of the rules by means of infinitesimal differentials appeared in Leibniz's Parisian manuscripts (see above), where higher-order differentials or products of differentials have to be discarded with respect to ordinary differentials, and similarly ordinary differentials have to be discarded with respect to finite quantities. Conscious of the criticism that the use of the infinitely small quantities would have had on the contemporaries, Leibniz chose to hide it in his first paper; many years later, replying to the objections of Bernard Nieuwentijt, he showed in a manuscript how to prove the rules of the calculus

without infinitesimals, based on a law of continuity [Gerhardt, 1846, 39–50; trans. Child, 1920, 147–155]. In his 'Nova methodus' of October 1684 he would then go onto studying the behaviour of the curve in an interval, specifically increasing or decreasing ordinates, maxima and minima, concavity and convexity referred to the axis, the inflexion point and deducing the properties of differentials.

Leibniz implicitly always considered dx to be increasing positively, corresponding to the way in which dv was analysed. If dv were positive, the tangent would meet the x-axis towards the origin of the co-ordinates. This happened when the ordinates increased with the abscissae; if dv were negative, then the tangent would be drawn on the opposite side, which happened when the ordinates decreased when the abscissas increased. He then reversed matters and erroneously affirmed that increasing ordinates corresponded to positive dv and tangent leading to origin while decreasing ordinates corresponded to negative dv and tangent traced on the opposite side.

At this point, Leibniz neglected the case in which the curve increased with a point of inflexion having tangent parallel to the x-axis, and also other cases he considered afterwards in his manuscripts. After observing that $dv = 0$ for maxima or minima and the tangent line was parallel to the axis, it appeared that the differential was to be set equal to zero to find the maxima or minima and that the condition of the differential equalling zero at a certain point would determine the presence of maxima or minima there. Leibniz did not question the existence of singular points, maybe because the matter was already evident from geometric or physical considerations, and he was only concerned about their exact determination. Only several years later did the idea dawn on him that the sign of the differential ratio (or of the derivative) changed from $+$ to $-$, which could occur either through zero or through infinite (as specified by Johann Bernoulli and Guillaume François de l'Hôpital (1661–1704) and published in the textbook *Analyse des infiniment petits* of 1696) or maintain the function finite-valued and not null (as later proved by Augustin-Louis Cauchy: §25) was fundamental in determination of maxima and minima.

After introducing the concept of convexity and concavity referred to the axis and linked to increase and decrease of ordinates and of the prime differentials, Leibniz dealt with the second differentials, simply called 'differences of differences' for which constant dx was implicitly presupposed. The inflexion point was thus defined as the point where concavity and convexity were exchanged or as a maximum or minimum of the prime differential. These considerations, burdened by the previous incorrect double implications, would lead him to state as necessary and sufficient conditions which were in fact only necessary. They will be elucidated in l'Hôpital's textbook of 1696.

Leibniz then set out the differentiation rules for powers, roots and composite functions. In the latter case, he chose to connect a generic curve to the cycloid because he wanted to demonstrate that his calculus was easily also applied to transcendent curves, possibility that Descartes wanted to exclude from geometry. It was a winning move to attract the attention on one of the most celebrated curves of the time, and his mentor Huygens expressed to him his admiration when in 1690 Leibniz sent him in detail the calculation of the tangent to the cycloid [Dupont and Roero, 1991, 117–119].

Finally, Leibniz demonstrated how to apply his differential method on four current problems which led him to proudly announce the phrase quoted at the beginning of this paper.

The first example, on the determination of a tangent to a curve, was very complex, containing many fractions and radicals. Earlier methods of past and contemporary mathematicians, such as Descartes, P. de Fermat, Jan Hudde and Sluse, would have required very long calculations. The second example was a minimum problem occurring in refraction of light studied by Descartes and by Fermat. Fermat's method for maxima and minima led to an equation containing four roots, and hence to long and tedious calculations [Andersen, 1983]. The third example was a problem that Descartes had put to Fermat, deeming it 'of insuperable difficulty' (Fermat to P. de Carcavi, 20 August 1650) because the equation of the curve whose tangent was to be determined contained four roots. Leibniz complicated the curve whose tangent was sought even more because his equation contained six. He solved a similar problem in a letter sent to Huygens on 8 September 1679 [Gerhardt, 1850, 17–38]. The last argument was the 'inverse-tangent problem', which corresponded to the solution of a differential equation, that is, find a curve such that for each point the subtangent is always equal to a given constant. In this case, the problem was put by Florimond de Beaune to Descartes, who did not manage to solve it, while Leibniz reached the goal in only a few steps. By these four examples he demonstrated the power of his differential method.

3 THE EARLY RECEPTION OF THE DIFFERENTIAL CALCULUS

The brevity of Leibniz's essay of October 1684 bewildered many readers: it was not immediately understood even by the most famous mathematicians. Jacob Bernoulli (1654–1705), who was to become one of the main supporters of the new calculus, was the first to underline these difficulties. In 1687, soon after obtaining the mathematics chair at University of Basel, he wrote to Leibniz on 15 December 1687 asking for explanations on the mysterious new method [Gerhardt, 1855, 13]:

> Therefore I believe that you, Monsieur, are concealing here the traces of a more sublime form of mathematics that I have not yet succeeded in penetrating using common Cartesian analysis. I wish to learn about the mathematics by means of which you and Messer Tschirnhaus have discovered so many and such important things on the squaring of the circle and on the dimensions of other curves. If you will deem me worthy of partaking a ray of light of your method (that I most dearly wish), to the extent allowed by your very important commitments, once I have been informed of your discoveries, I will become not only a simple admirer, but your most devoted appraiser and propagator.

Unfortunately, Leibniz was not able to answer this request for explanation because he was travelling to southern Germany, Austria and Italy at that time, in search of the origins of the House of Brunswick-Lüneburg. Bernoulli, who tenaciously persevered in his intent, finally succeeded in penetrating the secrets of calculus and taught them to his younger brother Johann (1667–1748). As Jacob wrote to Leibniz on 15 November 1702: 'Once, before writing a letter to you, to justify myself, I intended to write a short story of my life and the discoveries we made since our tender youth (where, by the way, you would have noted that I, not he, penetrated the mysteries of your calculus first and that I them taught

them to him)' [Gerhardt, 1855, 63]. Neither did he hide the difficulties encountered from the start to the readers of his earliest publications, stressing the need to show use 'so that, if by chance my readers have not adequately understood the thoughts of the sharp scientist from the words he published in *Acta* in 1684, due to excessive brevity, they may learn his method of application here' [Jacob Bernoulli, 1691, 13; 1744, 431].

Evidence along these lines was confirmed in the obituaries written in 1705 following Jacob's death. For example, Bernard Fontenelle wrote: 'He had already gained understanding of the most abstruse geometry and was perfecting it with his own discoveries gradually as his studies progressed when, in 1684, geometry changed appearance nearly suddenly. Mr. Leibniz had given a few examples of the differential or infinitesimal calculus in the Leipzig *Acta* concealing both the art and the method' [Fontenelle, 1706, 139].

Johann Bernoulli expressed even more severe opinions about Leibniz's *Nova methodus* in his inaugural dissertation [1705] at Basel University:

> By the incomparable Leibniz was invented that famous calculus called differential to which all the questions that go beyond the common algebra are submitted and the curves that Descartes excluded from geometry are treated and expressed by their equations. Nevertheless to that renowned scholar pleased to show to the mathematical community the beauty of his invention only through a fog curtain. Precisely in 1684 in *Acta Eruditorum* the monthly journal edited in Leipzig he wanted to display the first elements only in very few pages, without any explanation, but covered in enigma.

In his autobiography Bernoulli repeated: 'After this beginning, by pure chance, my brother and I run into a short essay by Mr. Leibniz in the 1684 Leipzig *Acta*, where in barely five or six pages he sketched a very vague idea of differential calculus, which was more an enigma than an explanation; however, this was sufficient for us to grasp the entire secret in a few days. Evidence of this is found in our later publications on infinitesimals' [Wolf, 1848, 218–219].

Leibniz himself must have soon come to know of the difficulties encountered by his readers; for two years later he was to start off his second paper, on his integral calculus, thus: 'I have seen that many essays I published in *Acta* on the progress of Geometry have been most appreciated by many a man of culture and have gradually even entered common use, but, due to errors made in writing and to other reasons, some points were not fully understood, therefore I believe it will be worth adding a few remarks to clarify the previous articles herein' [Gerhardt, 1858, 226]. Later, in 1693, he informed Malebranche: 'If one day I will have the opportunity, I will present the rules and use of this calculus a little more clearly than I did in the Leipzig *Acta* because many mistakes are responsible for making the writing obscure and for this reason I believe that many readers did not understand a thing' [Gerhardt, 1875, 349].

The misprints and the obscurity of the text were a hurdle even for Huygens, who had spoken to Leibniz in Paris and had the opportunity to appreciate his scientific qualities. As he candidly confessed in a letter dated 24 August 1690: 'I read something about your new algebraic calculus in the Leipzig *Acta* but I found it obscure and did not study it sufficiently to gain understanding' [Gerhardt, 1850, 45]. On 9 October 1690, he wrote: 'I attempted after the aforesaid letter to understand your differential calculus and I insisted

to the extent that I can now comprehend, if only after two days, your examples, the one on the cycloid in your letter and the other on the search for Mr. Fermat's Theorem in the 1684 Leipzig Journal. I grasped the fundaments of calculus and of your entire method, which I deem very good and very useful' [Gerhardt, 1850, 47].

Once again, in September 1693, Huygens congratulated Leibniz on the success of his new analysis but did not hesitate to ask for additional explanations: 'I increasingly appreciate the beauty of geometry, for the new progresses which are made every day, in which you always play such an important role, Sir, thanks to your marvellous calculus, if not more. I am inadequately studying the matter but still do not understand the ddx at all and I would like to know if you have encountered major problems where it should be used, because I wish to study it' [Gerhardt, 1850, 162].

Considering the difficulties in understanding Leibniz's essay encountered by famous mathematicians mentioned above, one can easily guess the impact of the text on the averagedly educated *Acta Eruditorum* reader. Criticisms by historians were to be no less pungent than those expressed by his contemporaries [Dupont and Roero, 1991, 10–14]. In his history of mathematics (§21) Etienne Montucla affirmed that still many years after the essay was published 'the differential and integral calculus remained a mystery for most geometers' [Montucla, 1799, 397]. Later, Joseph Hofmann was even more severe and in 1966 suggested that Leibniz may have been deliberately obscure: '[*Nova Methodus*] was deliberately formulated in such a concise way that an unsuspecting reader could scarcely, if ever, grasp the fundamental ideas underlying the symbolism' [Hofmann, 1966, 219]. Leibniz's obscurity was also stressed by historians of philosophy such as A. Rivaud, who was sceptical about the possibility of understanding other essays on Leibniz's calculus, in addition to *Nova Methodus* [Rivaud, 1950, 486]:

> The famous essay published in *Acta Eruditorum* in October 1684, called *Nova Methodus pro maximis et minimis*, which contains the principles of the method, is most obscure and difficult. Later essays in actual fact only educated but a few exceptionally gifted readers, like the Bernoulli brothers Jacob and Johann and the Marquis de l'Hôpital. Even great mathematicians like Christiaan Huygens would never full comprehend the method that Leibniz had described to him in short excerpt in 1680.

4 THE FIRST PAPERS ON INTEGRAL CALCULUS

It was the book by John Craig on quadratures, *Methodus figurarum lineis rectis et curvis comprehensarum quadraturas determinandi*, published in London in 1685, that stimulated Leibniz to draft his first exposition of the integral calculus. Craig attributed to him the paper [Tschirnhaus, 1683b] on figures, published in the *Acta Eruditorum*. To clarify any false impressions concerning his own methods, Leibniz therefore sent to Leipzig his first article on the integral calculus, our second paper here: entitled 'About the deep geometry and the Analysis of indivisibles and infinities', it appeared in June 1686 in the *Acta Eruditorum*. Here he illustrated the importance of transcendent curves in the problems of quadrature and claimed the merit for having subjected this type of curve to analysis. Having defined the differential expression of the subnormal of a given curve, he went on to

'summation', i.e. he introduced *calculus summatorius*, the name of which reflects the link with the 'sums' of small rectangles whose heights are the ordinates and whose bases are the infinitesimal differences of the abscissas. The integration was presented as the inverse operation of differentiation and would be later accepted as such. The integrals considered were not indefinite integrals: from the differential equation $p\,dy = x\,dx$ Leibniz deduced $\int p\,dy = \frac{1}{2}x^2$ to be interpreted as $\int_0^y p\,dy = \frac{1}{2}x^2$. As early as his Parisian period, Leibniz decided to indicate the integral sign $\int$ with the stylised symbol of the initial of the word 'summa' [Hofmann, 1974, 187–193]. The term 'integral' appeared in May 1690 in an article by Jacob Bernoulli, and was possibly due to his brother Johann who claimed priority in his autobiography written in French in the 1740s [Jacob Bernoulli, 1690, 218; 1744, 423; *Works*, vol. 5, 30].

The central problem of quadratures is dealt with in our third paper, 'Supplement to measuring geometry', which appeared in September 1693 in the *Acta Eruditorum*. Leibniz introduced and constructed by tractional motion the 'quadratrix curve', whose ordinates represent the areas under the given curve. He showed that all quadrature problems could be reduced to inverse tangent problems; more precisely, the slope of the quadratrix curve results from the function to be squared, i.e. $\int_0^y z(y)\,dy$ was simply the ordinate $x = x(y)$ of a curve such that $\frac{dx}{dy} = z$, which in modern terms is expressed by saying that function x is the primitive of function z. There was no explicit definition of definite integral and indefinite integral, for which the same notation is used, but Leibniz and the Bernoulli brothers knew that infinite primitives, which differed by a constant term, correspond to an assigned function and knew the rule for computing $\int_0^y z(y)\,dy$ as the difference of the values assumed by a primitive in the bounds of integration (see Johann Bernoulli to Pierre Varignon, 11 August 1696, in his [1988, 106; 1742, vol. 3, 412–413]).

5 THE SPREAD OF THE LEIBNIZIAN CALCULUS

From the first, when Leibniz was living in Paris, he had understood that the algorithm that he had invented was not merely important but revolutionary for mathematics as a whole. Although his first paper on differential calculus proved to be unpalatable for most of his readers, he had the good fortune to find champions like the Bernoulli brothers, and a populariser like de l'Hôpital, who helped to promote and advance his methods at the highest level. There was certainly no better publicity for the Leibnizian calculus than the results published in the *Acta Eruditorum*, and in the Memoirs of the Paris and Berlin Academies. They not only offered a final solution to open problems such as those of the catenary, the brachistochrone, the velary (the curve of the sail when moved by the wind), the paracentric isochrone, the elastica, and various isoperimetrical problems; they also provided tools for dealing with more general tasks, such as the solution of differential equations, the construction of transcendental curves, the integration of rational and irrational expressions, and the rectification of curves. Both the mathematicians and the scholars of applied disciplines such as optics, mechanics, architecture, acoustics, astronomy, hydraulics and medicine, were to find the Leibnizian methods useful, nimble and elegant as an aid in forming and solving their problems.

The first and most important champion of the spread of the Leibnizian calculus was Leibniz himself. From the 1690s he had been looking among his friends for young collab-

orators who would be capable of assisting him in the preparation of the 'Scientia infiniti', a work designed to offer a clear exposition of infinitesimal mathematical procedures [Costabel, 1968; Dupont and Roero, 1991, 15–19]; but this work never saw the light. Leibniz had expressed his satisfaction at the publication in 1696 of de l'Hôpital's *Analyse des infiniment petits*, and had repeatedly said that he shared his invention with the Bernoulli brothers because of their elegant, and very profound, applications and further developments of his calculus that had made a significant contribution to its acceptance among scholars in scientific fields ranging from geometry to physics, from astronomy to mechanics, from optics to medicine.

Thus Leibniz's words are significant. 'I have so many different kinds of occupations to attend to that I gladly entrust this terrain to the cultivation of my friends' (Leibniz to Magliabechi, 18 August 1692, in his [*Works*, ser. 1, vol. 8, 395]); or 'But it is a very great pleasure for me to see the seeds I have sown bear fruit in others' gardens too' [Gerhardt, 1858, 258]; or, finally, 'Moreover as I am often more inclined to provide inspiration for discoveries than to make them, *acting as a whetstone which, though it does not itself cut, is in the habit of making iron sharp*, I hope I may be allowed sometimes to present things related either to calculations and reasonings or to execution, and to let others judge whether they are worth pursuing (Leibniz to l'Hôpital, 13/23 March 1699, in [Gerhardt, 1850, 333]).

It was through his wide network of acquaintances in various European countries that Leibniz put into effect all his strategies for the spread of his analysis. The presence first of Jacob Hermann, the favourite pupil of Jacob Bernoulli, and then of Nicolaus I Bernoulli, the nephew of the Bernoulli brothers, as professors of mathematics in Padua was one outlet [Mazzone and Roero, 1997]. In France it was through the Oratorian circle of Nicolas Malebranche (1638–1715) that Johann Bernoulli introduced in 1691 the Leibnizian calculus. His lessons to the Marquis de l'Hôpital led to the draft of the first treatise of differential calculus (1696), and it was under the influence of Malebranche that some years later appeared the first works on the integral calculus by Louis Carré in 1700 and Charles René Reyneau in 1708. The spread and acceptance of the Leibnizian calculus was transferred in this way to the wide public, through the manuals and textbooks written for students at universities or ecclesiastical colleges.

What Leibniz and Jacob Bernoulli had foreseen at the very beginning of the history of differential calculus was at last happening. For example, 'These are only the premises of a more sublime geometry which extends to all other more difficult and most beautiful problems, also of mixed mathematics, which without our differential calculus or the like, no-one is capable of treating with equal ease' (*Nova methodus*, 473). Or again, 'Besides, for all these problems that some have attempted to solve with other methods with no avail, I am convinced that the excellent and extraordinary use of Leibnizian calculus is necessary, to the extent that I believe that it should therefore be numbered among the most important discoveries of our century' [Jacob Bernoulli, 1691, 290; 1744, 452–453].

But this success was soon smeared with tragedy. Leibniz's calculus became known in the Continent before Newton's, which was published only in 1704. By then Newton and some of his followers were convinced that Leibniz had created his calculus by plagiarism, and Newton himself rigged an international committee at the Royal Society to come to the same conclusion [Hall, 1980]. Leibniz's followers, especially Johann Bernoulli, reacted

strongly against this slander, with the result that the mathematical community became polarised: Newton's theory was practiced almost exclusively in Britain and Leibniz's on the Continent, the latter with eventual greater success.

BIBLIOGRAPHY

Acta Eruditorum. 1693. *Indices generales auctorum et rerum* . . ., Leipzig: Haeredes et Gleditschium.

Aiton, E.J. 1985. *Leibniz. A Biography*, Bristol and Boston: Hilger.

Andersen, K. 1983. 'The mathematical technique in Fermat's deduction of the law of refraction', *Historia mathematica, 10*, 48–62.

Bernoulli, Jacob. *Works*. *Die Werke von Jacob Bernoulli*, in progress, Basel: Birkhäuser, 1969–.

Bernoulli, Jacob. 1744. *Opera*, 2 vols., Geneva: Cramer & Philibert.

Bernoulli, Jacob. 1690. 'Analysis problematis ante hac propositi, de inventione lineae descensus . . .', *Acta Eruditorum*, 217–219. [Repr. in 1744, 421–426; also in *Works*, vol. 5 (1999), 28–31.]

Bernoulli, Jacob. 1691. 'Specimen calculi differentialis in dimensione parabolae helicoidis . . .', *Acta Eruditorum*, 13–23. [Repr. in 1744, 431–442; also in *Works*, vol. 5 (1999), 32–47.]

Bernoulli, Johann. 1705. 'De fato novae analyseos et profundioris geometriae', manuscript, Basel, 17 November 1705, Universitätsbibliothek Basel.

Bernoulli, Johann. 1742. *Opera omnia*, 4 vols., Geneva: Cramer & Philibert.

Bernoulli, Johann. 1988. *Der Briefwechsel von Johann Bernoulli*, vol. 2 (ed. P. Costabel and J. Peiffer), Basel: Birkhäuser.

Bos, H.J.M. 1974. 'Differentials, higher-order differentials and the derivative in the Leibnizian calculus', *Archive for history of exact sciences, 14*, 1–90.

Bos, H.J.M. 1980. 'Newton, Leibniz and the Leibnizian tradition', in I. Grattan-Guinness (ed.), *From the calculus to set theory, 1630–1910: an introductory history*, London: Duckworth, 49–93.

Child, J.M. 1920. *The early mathematical manuscripts of Leibniz*, Chicago: Open Court.

Costabel, P. 1968. 'De scientia infiniti', in *Leibniz 1646–1716. Aspects de l'homme et de l'oeuvre*, Paris : Centre International de Synthése, 105–117.

Dupont, P. and Roero, C.S. 1991. *Leibniz 84. Il decollo enigmatico del calcolo differenziale*, Rende (CS): Mediterranean Press.

Fontenelle, B. Le Bovier de. 1706. 'Eloge de [Jacques] Bernoulli', *Histoire de l'Académie Royale des Sciences*, (1705), 139–150.

Gerhardt, C.I. 1846. *Historia et origo calculi differentialis a G.G. Leibnitio conscripta*, Hannover: Hahn'sche Buchhandlung.

Gerhardt, C.I. 1855. *Die Entdeckung der höheren Analysis*, Halle: Schmidt.

Gerhardt, C.I. (ed.) 1849–1863. *Leibnizens mathematische Schriften*, vol. 1 (1849), vol. 2 (1850), vol. 3 (1855), vol. 4 (1859), vol. 5 (1858), vol. 6 (1860), vol. 7 (1863), Halle: Schmidt. [Repr. Hildesheim: Olms, 1971.]

Gerhardt, C.I. (ed.) 1875–1890. *Die philosophischen Schriften von G.W. Leibniz*, 7 vols., Berlin: Weidmann. [Repr. Hildesheim: Olms, 1965.]

Giusti E. 1988. 'Il calcolo infinitesimale tra Leibniz e Newton', *Rendiconti del Seminario Matematico del'Universitá Politecnico di Torino, 46*, 1–29. [Repr. in *Giornale di fisica, 31* (1990), 47–59.]

Hall, R. 1980. *Philosophers at war. The quarrel between Newton and Leibniz*, Cambridge: Cambridge University Press.

Heinekamp, A. (ed.) 1986. *300 Jahre 'Nova Methodus' von G.W. Leibniz (1684–1984)*, Stuttgart: Steiner (*Studia Leibnitiana*, Sonderheft 14).

Hess, H.-J. 1986. 'Zur Vorgeschichte der Nova Methodus (1676–1684)', in [Heinekamp, 1986], 64–102.

Hofmann, J.E. 1949. *Die Entwicklungsgeschichte der Leibnizschen Mathematik während des Aufenthalts in Paris (1672–1676)*, Munich: Oldenbourg.

Hofmann, J.E. 1966. 'Zum öffentlichen Bekanntwerden der Leibnizschen Infinitesimalmathematik', *Sitzungsberichte der österreichishen Akademie der Wissenschaften, mathematische-naturwissenschaftliche Klasse*, Abt. 2, *175*, 209–254.

Hofmann, J.E. 1974. *Leibniz in Paris (1672–1676)*, Cambridge: Cambridge University Press. [Trans. of [1949].]

Leibniz, G.W. *Works. Sämtliche Schriften und Briefe*, several series, in progress, Leipzig and Berlin: Akademie der Wissenschaften, 1923–.

Mazzone, S. and Roero, C.S. 1997. *Jacob Hermann and the diffusion of the Leibnizian calculus in Italy*, Florence: Olschki.

Montucla, J.F. 1799. *Histoire des mathématiques*, 2nd ed., vol. 2, Paris : Agasse. [Repr. Paris : Blanchard, 1968. See §21.]

Parmentier, M. 1989. *G.W. Leibniz La naissance du calcul différentiel*, Paris : Vrin.

Rivaud, A. 1950. *Histoire de la philosophie*, vol. 3, Paris : Presses Universitaires de France.

Roero, C.S. 2002. 'Diffusione e primi sviluppi del calcolo infinitesimale', in *Storia della scienza*, vol. 5, *La rivoluzione scientifica*, Roma: Istituto della Enciclopedia Italiana Treccani, 474–486.

Scriba C.J. 1962–1966. 'The inverse method of tangents: a dialogue between Leibniz and Newton (1675–1677)', *Archive for history of exact sciences, 2*, 112–137.

Tschirnhaus, E.W. von. 1682. 'Nova methodus tangentes curvarum expedite determinandi', *Acta Eruditorum*, 391–393.

Tschirnhaus, E.W. von. 1683a. 'Nova methodus determinandi maxima et minima', *Acta Eruditorum*, 122–124.

Tschirnhaus, E.W. von. 1683b. 'Methodus datae figurae, aut quadraturam, aut impossibilitatem ejusdem quadraturae determinandi', *Acta Eruditorum*, 433–437.

Wolf, R. 1848. 'Erinnerungen an Johann I Bernoulli aus Basel', *Mitteilungen der Naturforschende Gesellschaft zu Bern*, nos. 136–137, 217–228.

CHAPTER 5

ISAAC NEWTON, *PHILOSOPHIAE NATURALIS PRINCIPIA MATHEMATICA*, FIRST EDITION (1687)

Niccolò Guicciardini

Newton's *Principia* is one of the great classics of the Scientific Revolution. Before 1687 natural philosophers were able to mathematize only parabolic motion caused by a constant force and circular uniform motion. Newton was pushing exact quantitative mathematization in fields such as the attraction exerted by extended bodies, the perturbed motions of many bodies in gravitational interaction, the motion in resisting media. The book delivered an awesome picture of the world, a world in which the same physical law governs celestial and terrestrial phenomena.

First edition. London: for the Royal Society, about 5 July 1687. viii + 511 pages. Print-run: first issue about 250–350, second issue about 50 copies. [Available on the internet at www.bnf.fr/gallica.]

Second edition. Cambridge: Cambridge University Press, 11–14 July 1713. xxviii + 492 pages. Edited by Roger Cotes, with a Preface concerning the Newtonian method in natural philosophy (pp. xi–xxvi). Print-run: 711 copies. [Contains many variants, emendations and additions (most notably the concluding General Scholium).]

Third edition. London: for the Royal Society, presentation copies available by 31 March 1726. xxxii + 536 pages. Edited by Henry Pemberton. Print-run: first issue 1000 copies, second issue 200, third issue 50.

Variorum edition = [Newton, 1972]. *Philosophiae naturalis principia mathematica. The third edition (1726) with variant readings assembled and edited by Alexandre Koyré and I. Bernard Cohen, with the assistance of Anne Whitman*, Cambridge: Cambridge University Press, 1972. [The starting point for any serious research on the *Principia*, it consists of a facsimile of the third Latin edition. The critical notes indicate the variants relative to the first two editions, to the manuscript deposited for publication, to Newton's Lucasian lectures based on the *Principia*, as well as to Newton's annotations

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

and marginalia to several copies. Published along with [Cohen, 1971] which provides information on the history of the composition and diffusion of the text.]

Many reprints, abridgements, and translations into many languages including Chinese, Dutch, French, German, Italian, Japanese, Rumanian, Russian, Spanish and Swedish (for up to 1972 see [Newton, 1972, 855–883]). We note these:

Principal English translations. 1) *The mathematical principles of natural philosophy. By Sir Isaac Newton. Translated into English by Andrew Motte. To which are added the laws of the Moon's motion according to gravity. By John Machin*, London: for Benjamin Motte, 1729. [Photorepr. London: Dawson, 1968.] 2) This ed. revised and 'supplied with an historical and explanatory appendix' by Florian Cajori as *Sir Isaac Newton's mathematical principles of natural philosophy and his system of the world*, Berkeley, Los Angeles and London: University of California Press, 1934. [The best known translation; but corrigible, and now superseded by] 3) [Newton, 1999]. *The Principia: Mathematical principles of natural philosophy, A new translation by I. Bernard Cohen and Anne Whitman assisted by Julia Budenz, Preceded by a guide to Newton's Principia by I. Bernard Cohen*, Berkeley, Los Angeles and London: University of California Press, 1999. [Prefaced by an informative guide [Cohen, 1999], written by I.B. Cohen (with contributions by George Smith and Michael Nauenberg). 4) Partial translations in [Brackenridge, 1995] and [Densmore, 1995].]

German translation. *Die mathematischen Prinzipien der Physik* (ed. and trans. Volkmar Schüller), Berlin, New York: Walter de Gruyter, 1999. [Includes a translation of the variants noted in [Newton, 1972] and a critical apparatus.]

Spanish translations. 1) *Principios matemáticos de la filosofía natural* (trans. Antonio Escohotado), Madrid: Editora Nacional, 1982. [Repr. Madrid: Tecnos, 1987; and Barcelona: Altaya, 1994.] 2) *Principios matemáticos de la filosofía natural*, introduction, translation and notes by Eloy Rada García, Madrid: Alianza Editorial, 1987.

Manuscripts. The manuscript deposited for publication and the Lucasian lectures can be reconstructed from [Newton, 1972]. Manuscripts related to the *Principia* can be found especially in [Herivel, 1965]; Newton *Papers*, vol. 6; and *The preliminary manuscripts for Isaac Newton's 1687 Principia, 1684–1685: facsimile of the original autographs, now in Cambridge University Library* (ed. and introd. by D.T. Whiteside), Cambridge: Cambridge University Press, 1989.

Related articles: Descartes (§1), Leibniz (§4), Maclaurin (§10), d'Alembert (§11), Lagrange (§16, §19), Laplace on celestial mechanics (§18).

1 NEWTON'S MATHEMATICAL METHODS

The *Philosophiae naturalis principia mathematica* (hereafter the *Principia*) was published in 1687, thanks to the financial and editorial support of Edmond Halley (1656–1742), and under the auspices of the Royal Society. The author was the Lucasian Professor of Mathematics at Cambridge University. Since his election to the prestigious Chair in 1669, he had spent a rather monotonous life in Cambridge, interrupted by a few travels to his native

Lincolnshire and to London. He was known to his contemporaries as a talented mathematician and a competent divine, even though little of his mathematical discoveries and even less of his (heretical) theological studies had been communicated to the world. His theory of the refraction of light had indeed been printed from 1672 in the *Transactions* of the Royal Society, meeting skepticism and opposition, especially from the Society's curator of experiments, Robert Hooke (1635–1703).

1.1 Background

Isaac Newton (1643–1727) had entered Trinity College in 1661 while the University was paralized by the turmoil caused by the purges following the re-establishment of the Stuart monarchy. The confusion that reigned in those years allowed a studious lad to devote his time to reading and experimenting. Newton soon distanced himself from the standard curriculum, which assigned a rather dull diet of Aristotelianism, and began reading in the new geometry and natural philosophy. Boyle's corpuscularism and Descartes's theory of motion and cosmology attracted his attention. Somewhat dissatisfied by the materialism of the mechanical philosophy (an attempt to explain all the phenomena of nature in terms of the impact and shapes of material particles) he was soon to delve into the esoteric texts published during a renaissance of alchemical research which occurred in the first half of the 17th century. Newton flirted with alchemy almost all his life: a flirt which is still the cause of much disagreement amongst Newtonian scholars. His smoky 'elaboratory' was built just beside Trinity Chapel. Newton's last alchemical papers are dated 1696, when—now a celebrity (and from 1705 'Sir Isaac')—he moved to London to become rich as Warden (and then Master) of the Mint, and powerful in science as President of the Royal Society.

But Newton's first love was mathematics. Just a couple of years after his arrival on the Cam's shores he began reading some symbol-laden books which introduced him to the new symbolic algebra: William Oughtred's (1574–1660) *Clavis mathematicae* (Oxford, 1631) and Frans van Schooten's (1615–1660) *Exercitationes mathematicae* (Leiden, 1657). van Schooten, a well-known Dutch mathematician influenced by René Descartes (1596–1650), was the editor of François Viète's (1540–1603) collected works, the *Opera mathematica* (Leiden, 1646), which Newton studied. Newton was particularly impressed by two short, highly advanced, tracts: Descartes's *Géométrie* (1637) (§1) and John Wallis's (1616–1703) *Arithmetica infinitorum* (1656) (§2). The importance of Descartes's *Géométrie* for Newton's mathematical development cannot be overestimated. It is by reading the second Latin edition, translated by Frans van Schooten and enriched by commentaries by van Schooten himself and other Dutch mathematicians, that Newton learned how the study of plane curves could be carried on in algebraic terms. Newton devoted particular attention to two problems: the drawing of tangents to curves and the determination of the area subtended to a curve.

Newton immediately faced these two problems in terms which would have been inadmissible by the standards set in the *Géométrie*. Firstly, not only manipulation of equations, but also kinematic properties such as instantaneous velocity and infinitesimal displacements entered into the solution of geometric problems. Secondly, infinitesimals—conceived of as 'moments', infinitesimal increases covered in an infinitesimal interval of time—were admitted. Finally, the curves could be represented by 'infinite equations', viz.

infinite series, and not only by 'finite' algebraic equations. As a result of these departures from the *Géométrie* Newton could accept mechanical (i.e. in modern terms, 'transcendental') curves as admissible objects of mathematical inquiry. It is clear that Newton considered the exclusion of mechanical curves as a serious limitation of Descartes's method. He insisted that his method overcame such a limitation.

In a paper dated October 1665 entitled 'How to draw tangents to Mechanicall lines' [Newton *Papers*, vol. 1, 272–280] Newton conceived curves as the trajectory of a moving body. A possible source for Newton is Isaac Barrow (1630–1677), who conceived curves as generated by motion in very similar terms. The idea was, however, common in 17th-century mathematics, and an indirect influence of Gilles Personne de Roberval (1602–1675) should not be excluded. What is new with Newton is the fact that, thanks to the kinematic conception of geometric quantities, he arrived at the understanding of the inverse relationship between tangent- and area-problems (what in a somewhat Whiggish way we might call the 'fundamental theorem of the calculus'). Newton developed also an efficient algorithm, which can be roughly defined as 'equivalent' to Leibniz's differential and integral calculus (§4). Thanks to the fundamental theorem he began tackling quadrature problems by anti-differentiation.

1.2 Progress

In Winter 1664, inspired by Wallis's *Arithmetica infinitorum* (1656), thanks to a rather shaky inductive procedure, Newton stated the binomial theorem for fractional exponents. He immediately realised that quadrature problems (the inverse problems) could be tackled via infinite series: as we would say nowadays, by expanding the integrand in power series and integrating term-wise. Newton deployed this procedure from the very beginning of his mathematical researches. These infinitary problem-solving techniques formed the core of a treatise that Newton wrote in 1669 entitled 'De analysi per aequationes numero terminorum infinitas' (On the analysis by means of equations with an infinite number of terms) [Newton *Papers*, vol. 2, 206–247]. Equations with an infinite number of terms (i.e. infinite series) were not, of course, contemplated by Cartesian 'common analysis', or 'common algebra' (as Descartes's algorithm came to be known in the second half of the 17th century). Their use, according to Newton, allowed to extend the boundaries of analysis. Newton referred to his 'new analysis' as the 'method of series and fluxions'. He defined a 'fluent' as a magnitude which flows continuously in time (e.g. an area which increases by continuous motion of the ordinate). The 'fluxion' is the instantaneous velocity of the fluent.

In the 1670s Newton began to raise deliberate criticisms against Cartesian mathematics. As the time passed by, his distaste towards the 'analysis of the moderns' increased. This notwithstanding, Newton remained for all his life profoundly influenced by Cartesian mathematics, even though he would not have liked to admit it. The heritage of the *Géométrie* could not be obliterated and remained evident in works such as the *Arithmetica universalis* (1707) [Newton *Papers*, vol. 5, 54–491] and the *Enumeratio linearum tertii ordinis* (1704) [Newton *Papers*, vol. 7, 588–645]. However, from the early 1670s Newton began distancing himself from Cartesian mathematics, in order to devote his attention to the works of ancient geometers. He spent much effort to studying Pappus's *Collectio mathematica*.

Newton's new interest for geometry, an interest which he shared with many of his contemporaries such as Barrow and Thomas Hobbes, led him to compose several geometrical works which were unpublished until recently [*Papers*, vols. 4 and 7]. Now we know that Newton trod in the steps of those who believed that the ancients were possessors of a hidden analysis, a method of discovery superior to the analysis of the moderns. He rejected infinitesimals, a typical tool of modern mathematics, and tried to present his method of fluxions in purely geometric terms, avoiding the symbolism of algebra. Newton was led to distance himself from his early mathematical work on fluxions: he gave preference to a new method that he termed the 'synthetic method of fluxions', rather than to his earlier 'analytical method of fluxions'. In this new presentation of the fluxional method the importance of basing geometry upon kinematics was further enhanced.

The synthetic method of fluxions was first worked out in a treatise entitled *Geometria curvilinea*, written around 1680 [Newton *Papers*, vol. 4, 420–484]. Newton's purpose was to reformulate the results concerning fluents and fluxions, which he had achieved in his early analytical method of fluxions, in geometric terms compatible with the methods of the ancients. In the first place he had to avoid symbolic algebra: he did so by referring directly to geometric figures and their properties. Secondly he had to avoid infinitesimals: instead of making recourse to infinitesimals, he deployed limit procedures. We will find in section 4 many examples of these geometric limit procedures in the *Principia*.

Many reasons lay behind Newton's shift for the geometry of the ancients and his critical attitude towards the moderns, an attitude which is somewhat in resonance with his anti-cartesianism in philosophy, and with his belief in a *prisca philosophia* (the wisdom of the ancients) which he fully endorsed most probably in the 1690s. Here we note that he was often to underline the fact that the objects of the synthetic geometric method of fluxions have an existence in Nature. The objects of geometric inquiry are generated by motions which one observes in the study of the natural world (one can think of the ellipses traced by planets).

2 EARLY STUDIES ON THE MOTION OF BODIES AND ON PLANETARY MOTION

2.1 Initial influences

Newton's earliest studies on the laws of motions occurred during the Winter of 1664 [Herivel, 1965; Nauenberg, 1994]. As in the case of mathematics, his starting point was Descartes. He commented upon Book 2 of Descartes's *Principia philosophiae* (1644) with particular penetration. It is believed that the title of Newton's magnum opus was conceived of as a criticism to the French philosopher, whose work would have lacked adequate mathematical principles. From Descartes Newton learned about the law of inertia: what was to become the first law of motion of the *Principia*. A body moves in a straight line with constant speed until a force is applied to it. Unaccelerated rectilinear motion is a status in which a body naturally perseveres: it does not need, as it was thought in the Aristotelian tradition, a mover.

By the early 1660s natural philosophers had concerned themselves with two cases of accelerated motion: rectilinear uniformly accelerated and uniform circular motion. The

first case occurred in the fall of bodies and gave rise (as Galileo had taught) to parabolic trajectories by composition of inertial and uniformly accelerated motions. Uniform circular motion was often conceived of as caused by the balancing of two endeavours or conatuses: one centrifugal to recede from the centre, one centripetal seeking the centre. An additional endeavour acting at right angles to the radius was often posited in order to sustain the uniform circular motion.

From the very beginning of his studies Newton was trying to subject the motions of bodies to mathematical laws. His first mathematical law in this field is nowadays attributed to Christiaan Huygens (1629–1695) since it was first published in 1673 in the *Horologium oscillatorium* (§3). In modern terms the laws says that the centripetal acceleration of a body which moves in a circular trajectory with constant speed is proportional to the square of the speed and inversely proportional to the radius.

In 1665 Newton tried to generalise his mathematical results on uniform circular motion to more general cases. He noted [Whiteside, 1991, 14]:

> If the body b moved in an Ellipsis, then its force in each point (if its motion in that point bee given) may bee found by a tangent circle of Equall crookednesse with that point of the Ellipsis.

This is an extremely fertile insight. How can we go beyond the simple cases of uniformly accelerated rectilinear and circular uniform motion? In order to estimate the acceleration, for instance, in an elliptical trajectory (it goes without saying that here Newton had in mind planetary orbits), Newton conceived that locally the body moves with circular uniform motion along the osculating circle. It should be reminded that in those years he was developing fluxional techniques to calculate the radius of curvature to plane curves. Newton understood that the instantaneous normal acceleration a_N in a non circular orbit can be calculated by applying locally Huygens laws for circular uniform motion: in modern terms $|a_N| = v^2/\rho$ (v instantaneous speed, ρ radius of curvature) [Brackenridge, 1995].

What about Newton's early thoughts on planetary motions? The few extant records indicate that he remained for many years trapped in the framework of Cartesian vortex theory [Cohen, 1999, 11–22]. There are reasons to believe that in the 1660s Newton thought that the planets orbit the Sun because they are transferred by a vortex in nearly circular orbits. Assuming that the orbits are exactly circular, it was childplay, combining Huygens's law with Kepler's third law, to verify that the planets' radial acceleration varies inversely as the square of their distance from the Sun. This inverse square law was thus attained in a context which is far away from gravitation theory.

2.2 The role of Hooke

Instrumental in awakening Newton from his Cartesian dream was Hooke, who in 1679 tried to revive correspondence with an offended Newton because of the reception of the paper on the theory of light. Hooke, who had been recently appointed secretary of the Royal Society, proposed to Newton a new 'hypothesis' for planetary motion according to which planets, moving in vacuo, describe orbits around the Sun because of a rectilinear inertial motion by the tangent and an accelerated motion towards the Sun. Hooke was thus disposing both of the Cartesian endeavour to recede from the centre and of the Cartesian

vortex. Only one *centripetal* force directed towards the Sun is needed to deviate the planets unresisted inertial rectilinear motion. Newton was soon to discover that Hooke's hypothesis was mathematically fruitful. The most fruitful insight that Newton achieved, even though it is unclear exactly when he did so, was that a body moving in a space void of resistance and attracted by a central force must obey Kepler's area law, and vice versa a body which moves in accordance to the area law must be accelerated by a central force (more on this in section 5).

It is difficult to establish the steps which led Newton to conceive universal gravitation. Certainly Hooke's contribution was momentous, and historians are re-evaluating the role of the Royal Society's Secretary in formulating the new cosmology [Inwood, 2002; Bennett et alii, 2003; Nauenberg, 2004]. Hooke's ideas were indeed revolutionary, and the extant records prove that Newton did not immediately endorse them: he continued to believe that the planetary motions were caused by a revolving ether. Contrary to Descartes, he seems to have interpreted this ethereal medium in non-mechanistic terms, somewhat reminiscent of his alchemical researches. Until 1681 Newton discussed the motion of comets with John Flamsteed (1646–1719) in terms of a fluid which revolves round the centre of the cosmic system carrying with itself the planets and the comets [Ruffner, 2000]. Furthermore, he believed that the two appearances of the 1680 comet were actually due to two comets moving along roughly straight line trajectories. The appearance of the 1682 comet, whose trajectory passed close to the ecliptic but in the reverse direction of planetary orbits, probably gave the final blow, in Newton's mind, to the cosmic vortex. At last, Newton realised that the interplanetary space is void [Kollerstrom, 1999]. At least there is no inert matter there: Newton never abandoned the hypothesis of the existence of a non-material planetary medium completely.

Hooke's hypothesis for planetary motions was discussed at the Royal Society by astronomers interested in alternatives to Cartesian cosmology. Perhaps, as Kepler had suggested, the Sun was the cause of a force analogous to the force between loadstone and iron: this centripetal force would deviate the planets. But, how could one relate this force to the observed motions of the planets? More specifically, was it possible to prove any implication between the three Keplerian planetary laws and a specific force law? It was highly suspected that this law would be inverse square. Christopher Wren (1632–1723) posed this problem to Halley and Hooke. Could one of the two provide, within the context of Hooke's hypothesis, a mathematical theory linking the Keplerian laws with a specific force law? The winner would have been rewarded with a book worth 30 shillings.

Since finding a reply to Wren's question proved to be mathematically difficult, Halley took the wise, if somewhat humiliating, decision of travelling to Cambridge on August 1684 to ask Newton's advice. Most to his amazement he found that the Lucasian Professor had the answer, or, at least this is what he claimed. In November 1684 Halley received from Newton a short treatise, the *De motu*, in which Wren's desiderata were satisfied [De Gandt, 1995]. This is how the *Principia* began to take shape. Thanks to Halley's encouragement and insistence, a reluctant Newton was convinced to embark into a project that a couple of years afterwards—indeed years of hard work and scientific creativity—led to the completion of the *Principia*, the work to which now we turn.

3 THE *PRINCIPIA* (1687): DEFINITIONS AND LAWS

The *Principia* (2nd and 3rd eds.) opens with a long laudatory Ode written by Halley in honour of the author (Table 1). This is followed by several Prefaces of philosophical content, several pages on the basic definitions of mass ('quantity of matter'), momentum ('quantity of motion'), inherent, impressed, centripetal, absolute, accelerative, and motive force, a puzzling but profound scholium on absolute time and space, the three laws of motions and their corollaries. These introductory pages have quite rightly attracted the attention of scholars: in particular Newton's conceptions of absolute time and space [DiSalle, 2002]

Table 1. Summary of Preliminaries to *Principia*, 1st edition (1687) (title page + dedication + pp. i–iv + pp. 1–25).
The most notable variant is the addition of Cotes's Preface, which contains important philosophical considerations on the relationships between natural philosophy and religion, as well as a defence of Newton's concept of gravitation against the criticisms raised by Continental Cartesians.

1st edition	**Main variants**
Title page, 1687, for the Royal Society. The two issues have different title pages.	2nd ed.: 1713, Cambridge University Press. 3rd ed.: 1726, for the Royal Society.
Dedication to the Royal Society.	
Author's Preface: relations between rational mechanics and geometry, forces of nature discovered from phenomena of motions.	
Halley's Ode.	From 2nd ed. moved after dedication.
	Author's Preface to 2nd ed.
	Cotes's Preface (added in 2nd ed.) on Cartesian and Newtonian methods.
	Author's Preface to 3rd ed.
	Chapter Index added in 2nd ed.
Definitions of: (i) quantity of matter; (ii) quantity of motion; (iii) inherent force; (iv) impressed force; (v) centripetal force; (vi) absolute, (vii) accelerative, and (viii) motive quantity of centripetal force.	
Scholium on absolute time and space, rotating bucket.	
Axioms or laws of motions.	
Corollaries on composition of forces and inertial frames.	
Scholium on collisions and experimental demonstration of third law of motion.	

and of 'inherent force' have been widely discussed [Cohen, 2002]. It is worth quoting the three laws [Newton, 1999, 416–417]:

> Law 1: Every body perseveres in its state of being at rest or of moving uniformly straight forward, except insofar as it is compelled to change its state by forces impressed.
> Law 2: A change in motion is proportional to the motive force impressed and takes place along the straight line in which that force is impressed.
> Law 3: To every action there is always an opposite and equal reaction; in other words, the action of two bodies upon each other are always equal and always opposite in direction.

Numerous scholars have faced the question of the equivalence between Newton's second law quoted above and its modern formulation as $F = ma$. It should be noted that Newton's law is formulated as a proportion, not as an equation (as in the modern case). Further, in Newton's law no reference to time is made. According to some scholars Newton's second law is best explained as stating a proportionality between the intensity of an instantaneous impulse and a discontinuous change of momentum. This conception of an impulsive impressed force which causes discontinuous changes of momentum might be related to Newton's endorsement of atomism, where impacts between hard atoms cause instantaneous velocity changes. There is no doubt, however, that Newton understood and used also the continuous formulation of the second law, as it appears, for instance, from Prop. 24, Book 2, which contains a statement that a continuous force is proportional to the infinitesimal change of momentum acquired in an infinitesimal interval of time [Newton, 1999, 700].

4 THE *PRINCIPIA* (1687): LIMITS

The mathematically minded reader of the *Principia* will hastily skip these preliminaries before pausing at last on the first mathematical jewels. (Book 1 is summarised in Table 2.) They are contained in Section 1, Book 1, devoted to the method of first and ultimate ratios [De Gandt, 1995; Brackenridge, 1995]. In this Section we read a clear statement about the use of infinitesimals [Newton, 1999, 441–442]:

> whenever in what follows I consider quantities as consisting of particles or whenever I use curved line-elements in place of straight lines, I wish it always to be understood that I have in mind not indivisibles but evanescent divisibles, and not sums and ratios of definite parts but the limits of such sums and ratios, and that the force of such proofs always rests on the method of the preceding lemmas.

Newton points out that the method of first and ultimate ratios rests on the following Lemma 1 (p. 433):

> Quantities, and also ratios of quantities, which in any finite time constantly tend to equality, and which before the end of that time approach so close to one another that their difference is less than any given quantity, become ultimately equal.

Table 2. Summary by sections of Book 1 of *Principia*, 1st edition (1687) (pp. 26–235). One of the most notable variants concerns Prop. 6, Section 2, which is foundational for the treatment of central force motion. In the 2nd ed. Newton presents a new measure of central force. Propositions 7–13 in Sections 2 and 3, which concern central forces, are augmented by additional demonstrations in which the new measure of force is deployed. Also Cor. 1, Prop. 13, was significantly altered. For a detailed analysis of variants in Sections 2 and 3 see [Brackenridge, 1995].

	Title: On the Motion of Bodies, First Book	
Sect.	**1st edition**	**Main variants**
1	First and ultimate ratios: 11 Lemmas on geometric limit procedures as foundation of the whole work.	
2	Props. 1–2: Kepler's area law valid iff force central. Prop. 4: on circular uniform motion: $a = v^2/\rho$. Prop. 6: geometrical measure of central force. Props. 7–10: central force determined given the orbit.	 Altered in 2nd ed. New measure of central force based on radius of curvature added in 2nd ed. Additional proofs based on new Prop. 6 added in 2nd ed.
3	Props. 11–13: $1/r^2$ force deduced for Keplerian orbits. Cor. 1, Prop. 13 on inverse problem of central forces: just a statement.	Additional proofs based on new Prop. 6 added in 2nd ed. Expanded in 2nd and 3rd eds. Newton claims it is a sketch of a proof.
4–5	Geometry of conics, Pappus problem solved, projective transformations.	Plans in the 1690s to move these sections at the end as a separate treatise on ancient geometry.
6	Algebraical nonintegrability of ovals, Kepler problem, Newton–Raphson method.	Scholium on approximation of Kepler equation's roots reworked in 2nd ed.
7–8	Orbit found when central force given. Cor. 3, prop. 41: result obtained by integration.	
9	Precession of nearly circular orbits. Prop. 45: use of infinite series.	In Cor 2, Prop. 45, 3rd ed., Newton notes failure in accounting for Moon's apses motion: 'The advance of the apsis of the moon is about twice as swift' [Newton, 1999, 545].
10	Cycloidal pendular motion: extension of results obtained in Huygens's *Horologium oscillatorium*. Results obtained by integration techniques.	

Table 2. (*Continued*)

	Title: On the Motion of Bodies, First Book	
Sect.	**1st edition**	**Main variants**
11	Three-body problem: qualitative treatment of Moon's inequalities, tidal motion, precession of equinoxes. Some of these results dealt numerically in Book 3.	Lettering of figures changed in 2nd ed.
12–13	Attraction of spherical and non-spherical bodies. Some results obtained by integrations (esp. Cor. 2, Prop. 91). Scholium, Prop. 93: binomial theorem.	
14	Motion of small corpuscles: mathematization of corpuscular optics.	

Newton's *ad absurdum* proof runs as follows (p. 433):

> If you deny this, let them become ultimately unequal, and let their ultimate difference be D. Then they cannot approach so close to equality that their difference is less than the given difference D, contrary to the hypothesis.

This principle might be regarded as an anticipation of A.L. Cauchy's theory of limits (§25), but this would certainly be a mistake, since Newton's theory of limits is referred to a geometrical rather than a numerical model. The objects to which Newton applies his 'synthetic method of fluxions' or 'method of first and ultimate ratios' are geometrical quantities generated by continuous flow. A typical mathematical problem which occurs in the *Principia* is the study of the limit to which the ratio of two geometrical fluents tends when they simultaneously vanish (Newton uses the expression the 'limit of the ratio of two vanishing quantities').

Notice that Newton is careful in dealing always with ratios of 'vanishing quantities' which have well-defined limits. For instance, in Lemma 7 he shows that given a curve (Figure 1) 'the ultimate ratio of the arc [*ACB*], the chord [*AB*], and the tangent [*AD*] to one another' tends to 1 [Newton, 1999, 436].

In Lemma 2 Newton shows that the area of a curvilinear surface *AabcdE* (Figure 2) can be approached as the limit of the areas of rectilinear surfaces, inscribed *AKbLcMdD* or circumscribed *AalbmcndoE*. Each rectilinear surface is composed of a finite number of rectangles with equal bases *AB*, *BC*, *CD*, etc. The proof is magisterial in its simplicity. Its structure is still retained in present-day calculus textbooks in the definition of the Riemann integral. It consists in showing that the difference between the areas of the circumscribed and the inscribed figures tends to zero, as the number of rectangles is 'increased indefinitely'. In fact this difference is equal to the area of rectangle *ABla* which, 'because its width *AB* is diminished indefinitely, becomes less than any given rectangle' (p. 433).

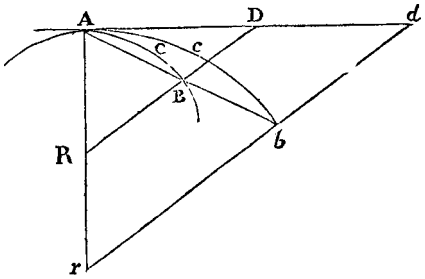


Figure 1. Limiting ratio of chord, tangent and arc (after Newton [1972, 79]).

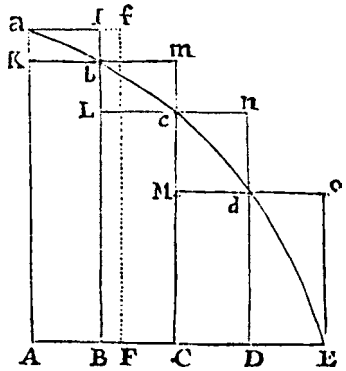


Figure 2. Approximating areas curvilinear surfaces (after Newton [1972, 74]).

Notice how in Lemmas 2 and 7 Newton provides proofs of two assumptions that were made in the 17th-century 'new analysis'. The 'new analysts' (Newton himself in his early writings!) had assumed that a curve can be conceived as a polygonal of infinitely many infinitesimal sides, and that a curvilinear surface can be conceived as composed of infinitely many infinitesimal stripes. According to Newton, the method of first and ultimate ratios provides a foundation for such infinitesimal procedures. In the *Geometria curvilinea* and in the *Principia* curves are smooth, and curvilinear surfaces are not resolved into infinitesimal elements. In the synthetic method of fluxions one always works with finite quantities and limits of ratios and sums of finite quantities.

Since Newton has banished infinitesimals and moments from the *Principia* in favour of limits, he has to justify the limits themselves, and in order to do so he makes use of geometrical and kinematical intuition. It is worth quoting from Section 1 at some length on this particular point (p. 442):

> It may be objected that there is no such thing as an ultimate proportion of vanishing quantities, inasmuch as before vanishing the proportion is not ultimate, and after vanishing it does not exist at all. But by the same argument it could equally be contended that there is no ultimate velocity of a body reaching a certain place at which the motion ceases; for before the body arrives at this place, the velocity is not the ultimate velocity, and when it arrives there, there is no velocity at all. But the answer is easy; to understand the ultimate velocity as that with which a body is moving, neither before it arrives at its ultimate place and the motion ceases, nor after it has arrived there, but at the very instant when it arrives, that is, the very velocity with which the body arrives at its ultimate place and with which the motion ceases. And similarly the ultimate ratio of vanishing quantities is to be understood not as the ratio of quantities before they vanish or after they have vanished, but the ratio with which they vanish.

We turn now to Lemma 9 (Figure 3) which states (p. 437):

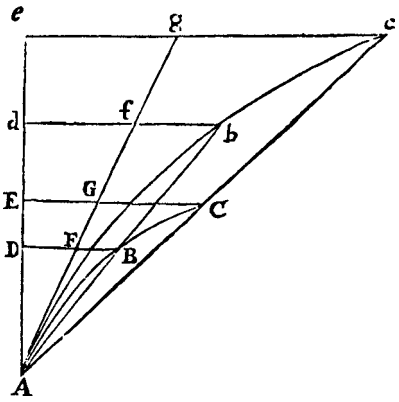


Figure 3. Lemma 9 (after Newton [1972, 80]).

> If the straight line *AE* and the curve *ABC*, both given in position, intersect each other at a given angle *A*, and if *BD* and *CE* are drawn as ordinates to the straight line *AE* at another given angle and meet the curve in *B* and *C*, and if then points *B* and *C* simultaneously approach point *A*, I say that the areas of the triangles *ABD* and *ACE* will ultimately be to each other as the squares of the sides.

As in Lemma 7, the essential step in the proof of Lemma 9 consists in local linearization. The curve *ABC* can be identified in the neighbourhood of *A* with its tangent *AG*. Therefore, if one takes the point *B* close to *A*, the curvilinear area *ABD* subtended by the curve grows very nearly as the square of the 'side' *AD*.

Lemma 9 has very important consequences for Newton's science of motion. These consequences are spelled out in Lemma 10. Let the abscissa *AD* in Figure 3 represent time, and the ordinate *DB* velocity, then (pp. 437–438):

> The spaces which a body describes when urged by any finite force, whether that force is determinate and immutable or is continually increased or continually decreased, are at the very beginning of the motion in the squared ratio of the times.

From Galileo's writings one knew that, when the force is 'determined and immutable', the space travelled from rest is proportional to the square of time (in Newton's terminology 'the spaces described are in the squared ratio of the times'). Newton states that this result is applicable to variable forces 'at the very beginning of the motion'. As Newton explains in the Corollaries to this Lemma, the first or ultimate ratio of displacement from rest, or generally displacement from inertial motion, over the square of the time is proportional to instantaneous acceleration.

As we will see, in the *Principia* Newton deals with variable forces. Thus he needs a geometric representation of force. Lemma 10 says that locally the velocity can be considered as varying linearly with time. Under this local approximation one can state that the displacement from inertial motion is proportional to the force times the square of time.

5 THE *PRINCIPIA* (1687): THE AREA LAW

One of the deepest dynamical insights that Newton gained after adopting Hooke's hypothesis on planetary motions is that Kepler area law is equivalent to central force motion. This is spelled out in Propositions 1 and 2, Book 1. These propositions are mathematically important since—when central force motion is considered—they allow Newton to represent geometrically time as the area swept by the radius vector, an essential step for achieving a geometrical representation of central force.

Proposition 1 reads as follows (p. 444):

> The areas which bodies made to move in orbits describe by radii drawn to an unmoving centre of forces lie in unmoving planes and are proportional to the times.

A body is fired at A with given initial velocity in the direction AB (Figure 4). The centripetal force acting on the body must be first imagined as consisting of a series of impulses which act after equal finite intervals of time. The trajectory will then be a polygonal $ABCDEF$. The body moves, during the first interval of time, from A to B with uniform rectilinear inertial motion. If the impulse did not act at B the body would continue its rectilinear uniform motion: it would reach c at the end of the second interval of time, so that $AB = Bc$. But because of the first impulse the body will be instantaneously deflected: it will reach C at the end of the second interval of time. Applying the first two laws of motion and elementary geometry, it is possible to show that triangles SAB and SBC have the same area and lie on the same plane. Similarly, all the triangular areas SCD, SDE, SEF, etc. spanned by the radius vector in equal times are equal and planar. In order to prove Proposition 1 Newton takes a limit. When the time interval tends to zero, the impulsive force approaches a continuous centripetal force and the trajectory approaches a smooth plane curve. The result (equal planar areas spanned in equal times by the radius vector)

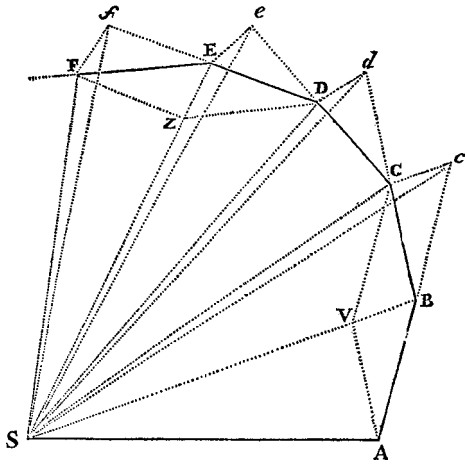


Figure 4. Polygonal trajectory (after Newton [1972, 90]).

obtained for the polygonal trajectory generated by the impulsive force is extrapolated to the limiting smooth trajectory generated by the continuous force.

Following a similar procedure, Newton proves Proposition 2 which states the inverse of Proposition 1. In Propositions 1 and 2 he has shown that a force is central if and only if the area law holds: the plane of orbital motion is constant and the radius vector sweeps equal areas in equal times. Notice that in his proof of Proposition 1 Newton makes recourse to limit arguments, according to the method of first and ultimate ratios.

6 THE *PRINCIPIA* (1687): CENTRAL FORCES

In order to tackle central forces with geometrical methods, a geometrical representation of such forces is required, a result which is not so easy to achieve since the central force applied to an orbiting body changes continuously, both in strength and direction. In Proposition 6 such a representation is provided.

This proposition implements Hooke's hypothesis. The body is accelerated in vacuo by a central force and its motion, as Hooke had suggested, is decomposed into an inertial motion along the tangent and an accelerated motion towards the force centre. A body accelerated by a centripetal force directed towards S (the centre of force) describes a trajectory as shown schematically in Figure 5. PQ is the arc traversed in a finite interval of time. The point Q is fluid in its position on the orbit, and one has to consider the limiting situation 'when points Q and P come together'. The line ZPR is the tangent to the orbit at P. QR tends to be parallel to SP as Q approaches P. QT is normal to SP. As we know from Lemma 10 (see Section 4), 'at the very beginning of the motion' the force can be considered as constant. In the case taken into consideration in Figure 5, this implies that, *as Q approaches P, the displacement QR is proportional to force times the square of time*. In fact, in the limiting situation, QR can be considered as a small Galilean fall caused by a constant force. Newton can now obtain the required geometrical representation of force. Since Kepler's area law holds (the force is central; cf. Prop. 1), the area of SPQ (a triangle since the limit of the ratio between the vanishing chord PQ and arc PQ is 1; cf. Lemma 7) is proportional to time. The area of triangle SPQ is $\frac{1}{2}(SP \cdot QT)$. Therefore, the geometrical measure of force is:

$$F \propto \frac{QR}{(SP \cdot QT)^2}, \tag{1}$$

where the above ratio has to be evaluated in the limiting situation 'when points P and Q come together' and $\propto$ is here used to mean 'is proportional to'. Proposition 6 is a good

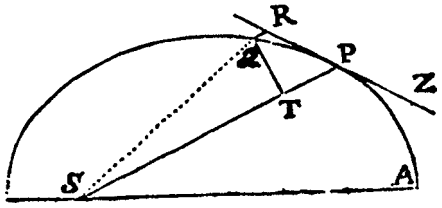


Figure 5. Parabolic trajectory in Prop. 6. After Newton [1972, 103n7].

example of application of the method of first and ultimate ratios. The limit to which tends the ratio $QR/(SP \cdot QT)^2$ is to be evaluated by purely geometric means. Notice that SP remains constant.

When the orbit is an equiangular spiral (in polar coordinates $\ln r = a\theta$) and S is placed at the centre (Proposition 9), as Q tends to P:

$$QR/QT^2 \propto 1/SP, \tag{2}$$

and thus the force varies inversely with the cube of distance. While, when the orbit is an ellipse and the centre of force is at the centre of the ellipse (Proposition 10) $QR/QT^2 \propto SP^3$: i.e. the force varies directly with distance. It is to be noted that forces which vary with the inverse of the cube of distance found an application later on in Book 3 in the study of tidal forces, while forces which vary directly with distance occur in the study of elastic vibrations.

In Section 3 Newton considers Keplerian orbits. In Proposition 11 Newton derives the result that if the body describes an orbit O, O is an ellipse and the force is directed toward a focus S, then the force varies inversely with the square of distance. In Propositions 12 and 13 he shows that the force is inverse square also if O is an hyperbola or a parabola. In fact, when the orbit is a conic section and S is placed at one focus,

$$QR/QT^2 \propto 1/L, \tag{3}$$

where L is a constant (the *latus rectum*), and the ratio QR/QT^2 is evaluated, as always, as the first or last ratio 'with the points Q and P coming together'. Therefore, for a body which obeys the first two Keplerian laws, the force varies inversely with the square of distance. This is the birth of gravitation theory [Brackenridge, 1995].

In Corollary 1 to Propositions 11–13 Newton states that if the force is inverse square, than the orbits are conic sections such that a focus coincides with the force centre. Corollary 1 reads as follows [Newton, 1999, 467]:

> From the last three propositions [i.e. Propositions 11–13] it follows that if any body P departs from the place P along any straight line PR with any velocity whatever and is at the same time acted upon by a centripetal force that is inversely proportional to the square of the distance from the centre, this body will move in some one of the conics having a focus in the centre of forces; and conversely.

Quite understandably this terse statement was subject to criticisms. Newton himself, while revising the second edition of the *Principia*, emended Corollary 1 adding the sketch of what he considered as a valid proof.

In the emended Corollary, and in the related Proposition 17, Newton makes it clear the following. He considers a 'body' of given mass m fired at P in an inverse-square force field. Initial velocity $\vec{v}_0$ (not directed towards the force centre) and normal component of central force $\vec{F}_{\mathrm{N}}$ at P determine the following geometric properties: i) the tangent at P, and ii) the curvature ρ^{-1} of the orbit at P (via local application of Huygens law, $\rho = mv_0^2/|F_{\mathrm{N}}|$). It is clear that Newton restricts his attention to a particular class of orbits, namely conic

sections having a focus at the force centre: let us call this class C. There is a unique conic with a focus in the force centre which satisfies the geometric properties implied by the initial conditions. Further, the initial conditions determine the areal velocity and thus a unique motion along this conic. From Propositions 11–13 one knows that such a conic motion is a possible trajectory. So what Newton does in Corollary 1 plus Proposition 17 is to show that for every initial condition it is possible to identify a conic belonging to C and a motion along it which satisfy the equation of motion. The proof that conics are necessary orbits is completed by adding a uniqueness condition (unproven by Newton) according to which for any initial condition only one trajectory is possible in a inverse-square force field. This Newtonian procedure has recently given rise to a lively debate [Pourciau, 1991; Weinstock, 2000]. This Corollary is an example of what was called 'an inverse problem of central forces': the central force F (force law and force centre S) is given, and what is required is the trajectory (a singular, assuming uniqueness!) corresponding to any initial position and velocity of a 'body' of given mass acted upon by such force.

7 THE *PRINCIPIA* (1687): PAPPUS'S PROBLEM

Descartes did not stress continuity with past tradition: his *Géométrie* could be read as a deliberate proof of the superiority of the new analytical methods, which united symbolic algebra and geometry, over those purely geometrical of the ancients. Descartes began the *Géométrie* with a problem, the four-lines locus, stated in Pappus's *Collectiones*. According to Descartes it could be inferred from Pappus's text, which he cited at length, that Euclid and Apollonius were not able to solve this problem, at least in its general form (§1.6). The so-called Pappus problem received a general solution in the *Géométrie*: could have been found a better proof of the superiority of the moderns over the ancients? In the late 1670s Newton studied in depth the seventh book of Pappus's *Collectiones* and began working on the restoration of some lost books by Euclid and Apollonius. These new interests led Newton to re-evaluate the use of geometry. Contrary to what Descartes had maintained, Newton believed that the solution of Pappus problem was within the reach of the ancients. His geometric solution of this problem attained not by 'a computation [as Descartes had done] but a geometrical synthesis, such as the ancients required' [Newton, 1999, 485] was to appear in print in Corollary 2 to Lemma 19, Section 5, Book 1, of the *Principia*. Sections 4 and 5 remain almost wholly separate from the rest of the *Principia*. They do not play a major role in Newton's mathematization of natural philosophy. In these Sections Newton aims at showing that the four-line locus can be determined by pure geometry. These two Sections are thus an anti-Cartesian manifesto [Di Sieno and Galuzzi, 1989].

In the 7th book of the *Collectiones* Pappus had hinted at the existence of a hidden method of discovery practiced by Euclid and Apollonius, the method of 'analysis'. Newton was convinced, by reading Pappus, that the ancient method of analysis was superior to the modern symbolic one, that he identified with Cartesian analysis. In his attempts to rediscover the method of the ancients Newton developed elements of projective geometry. The idea was that the ancients were able to solve complex problems related to conics by projective transformations (see Lemma 22). In Section 5 Newton applies projective transformations in order to determine the conic which is tangent to $5 - n$ given lines and touches n given points ($0 \leqslant n \leqslant 5$).

8 THE *PRINCIPIA* (1687): ALGEBRAIC NON-INTEGRABILITY OF OVALS

Section 6, Book 1, of the *Principia* is devoted to the solution of the so-called Kepler problem. The problem consists in finding the area of a focal sector of the ellipse and is equivalent to the solution for x of the equation $x - e \sin x = z$ (e and z given). Johannes Kepler (1571–1630) found that planets move in ellipses having the Sun placed at one focus. He also found that they move in such a way that the radius vector joining the Sun with the planet sweeps equal areas in equal times. When the elliptic orbit is known, the position of the planet in function of time can thus be found by calculating the area of the focal sector.

Lemma 28 reads as follows [Newton, 1999, 511]:

> No oval figure exists whose area, cut off by straight lines at will, can in general be found by means of equations finite in the number of their terms and dimensions.

This lemma is quite general. Given an oval figure and a point P inside it, we cut a sector S via a straight line passing through P. The lemma states that the sector S is not generally expressible by means of a finite algebraic equation. Peter Pesic has beautifully paraphrased Newton's simple, but powerful, demonstration [2001, 215]:

> Pick any point inside the oval and let it be the pole about which a line revolves with uniform angular speed. On that line, let a point move away from the pole with speed proportionate to the square of the distance along the line between the pole and the line's intersection with the oval. Then that moving point on the moving line will move in a gyrating spiral, its distance from the pole recording the area swept out by the line. The area of the oval is given by the distance moved by the point over one complete revolution of the line. But as the line continues to sweep over the oval area again and again, the spiral will continue uncoiling to infinity. Hence, it will intersect any straight line drawn across it an infinite number of times, which shows that the degree of the equation of the spiral is not finite, since an equation of finite degree can only intersect a given line a finite number of times. Therefore, since the area is given by the equation of the spiral, the area of the curve is not given by an equation of finite degree.

Newton's lemma 28 creates problems of interpretation since it is unclear what Newton means by 'an oval figure' [Pourciau, 2001]. However, one can assume that the ellipse is an oval figure and the Lemma applies to the focal sector of the ellipse swept by the radius vector according to Kepler's first two laws. Thus the Kepler problem can be solved only via infinite series (equations with an infinite number of terms). Infinite series, as we know, are Newton's main tool to deal with mechanical curves. In the following Proposition 31 and its scholium Newton shows how one can obtain numerical approximations of the equation $x - e \sin x = z$ thanks to an iterative procedure which is related to the Newton–Raphson method [Kollerstrom, 1992]. As Newton proves, Descartes's rejection of infinitary techniques would render the mathematization of Keplerian planetary theory an impossible task.

9 THE *PRINCIPIA* (1687): THE GENERAL INVERSE PROBLEM OF CENTRAL FORCES

In Sections 7 and 8 of Book 1 Newton faces the general inverse problem of central forces: i.e. the problem of determining the trajectory, given initial position and velocity, of a 'body acted upon' by *any* central force. He deals first with 'rectilinear ascent and descent' and then with curvilinear motion. The inverse problem for inverse square central forces was faced by Newton in Corollary 1 to Propositions 11–13 and in Proposition 17, Book 1 (Section 6). In Proposition 41 a general solution of the inverse problem is provided [Cohen, 1999, 334–345]. This proposition is based on the assumption that a method for the 'quadrature of curvilinear figures' is given. As we know, Newton had developed in his youth the 'method of fluxions and infinite series'. However, in the *Principia* he chose not to make his mathematical discoveries in this field wholly explicit. When Newton in the *Principia* reduces a problem to a difficult quadrature, he follows the policy of giving the solution without the complete demonstration. He simply shows that the solution depends upon the quadrature of a curve and leaves the reader without any hint on how to perform the required quadrature. Other examples of these mysterious reductions to quadratures can be found in Newton's treatment of the attraction of extended bodies (Sections 12 and 13, Book 1), of the solid of least resistance (Scholium to Proposition 35, Book 2), and of the inequalities of the Moon's motion (Propositions 26–35, Book 3). These parts of the *Principia* were really puzzling for his readers. They were told that a result depended upon the quadrature of a certain curve, but the method by which this very quadrature could be achieved was not revealed [Guicciardini, 1999].

As far as Proposition 41 is concerned the following points should be noticed: i) it is entirely 'geometrical', but it can be easily translated into calculus terms; ii) the final result is easily translatable into a couple of fluxional (or differential) equations; iii) Newton was aware that a translation into calculus was feasible.

The last point is supported by overwhelming evidence [Brackenridge, 2003]. Firstly, Newton writes that the demonstration of Proposition 41 depends upon the 'quadrature of curvilinear figures'. Secondly, in Corollary 3 he applies the general result of Proposition 41 to the case of an inverse cube force. The question to be answered in this Corollary is: which trajectories are described by a body accelerated by an inverse cube force? In Corollary 3 Newton gives only the solution in the form of a geometrical construction: he constructs some spiral trajectories which answer the problem. He could have obtained this result only by the quadrature (in Leibnizian terms, 'integration') of the fluxional (in Leibnizian terms, 'differential') equations expressed geometrically in Proposition 41. In Corollary 3 Newton does not perform this quadrature explicitly, but simply states the result. He then adds: 'All this follows from the foregoing proposition [41], by means of the quadrature of a certain curve, the finding of which, as being easy enough, I omit for the sake of brevity' [Newton, 1999, 532]. Thirdly, when David Gregory, during a visit he paid to Newton in May 1694, asked about the mysterious method applied in the solution of Corollary 3, the Lucasian Professor answered by translating the basic result of Proposition 41 as a fluxional equation. He applied this equation to the case of an inverse cube force and obtained the result stated in the *Principia*. As Gregory remarked in a memorandum of this visit, 'on these [quadratures] depend certain more abstruse parts in his philosophy as hitherto published,

such as Corollary 3, Proposition 41' [Newton *Correspondence*, vol. 3, 386]. There is no evidence that Newton ever applied Proposition 41 to inverse-square forces. The first ones to do so in print were Jacob Hermann and Johann Bernoulli in 1710.

10 THE *PRINCIPIA* (1687): UNIVERSAL GRAVITATION

In order to approach universal gravitation in mathematical terms Newton had to advance into unknown territory. Until Section 8, Book 1, he had dealt with a body moving in a central force field. Newton knew that this mathematical model can be applied only approximately to the planetary system. In practice when one considers a system composed of two bodies 1 and 2, sufficiently far from other disturbing bodies and sufficiently far one from the other, and 1 has a much greater mass than 2, then one can approximate 1 as an immovable centre of force and 2 as a point mass. This simplified model occurs also in Section 9, Book 1, devoted to the motion of the line of apsides. It is only in Section 11 that Newton considers the motion of two, or more than two, bodies which mutually attract each other; and only in Sections 12 and 13 that he pays attention to the shape of the bodies, and to the gravitational force exerted by such bodies. These concluding more advanced sections of Book 1 contain a wealth of results, especially on embryonic perturbation theory [Nauenberg, 2001; Wilson, 2001].

Book 2 (Table 3) is devoted to the motion of bodies in resisting media. It is rich in mathematical results: most notably, in the Scholium to Proposition 35 (= 34, in the 2nd ed.) Newton inaugurates variational methods by tackling the problem of the solid of least resistance. The concluding Section 9 leads to what Newton conceived of as a refutation of the vortex theory of planetary motions. Book 2 contains many pages devoted to experimental results on resisted motion. The mathematical parts of Book 2 are very problematic. Compared with the mathematical methods of the first Book, those of the second were considered, since Newton's times, the less satisfactory, and in some cases just mistaken.

In Book 3 (Tables 4 and 5) Newton applied the mathematical results achieved in the first Book to astronomy. In a sequence of opening propositions he was able to infer from astronomical data that the planetary motions are caused by a gravitational force. This force acts instantaneously, in void, and attracts two given masses with a strength proportional to the product of the masses and inversely proportional to the square of their distance. In the remaining part of Book 3, Newton, assuming the existence of such a force between any two masses in the whole universe, was able to give quantitative estimates of diverse phenomena such as the motion of the tides, the shape of the Earth, some of the inequalities of the Moon's motion, and the precession of equinoxes and the trajectories of comets. In Lemma 5, in dealing with cometary paths, Newton presents a method of interpolation which was to inspire researches by mathematicians such as James Stirling, Friedrich Wilhelm Bessel and Carl Friedrich Gauss.

11 REVISIONS (1690S), SECOND (1713) AND THIRD (1726) EDITIONS

During the early 1690s Newton considered radical restructurings of the *Principia*. Despite the fact that nothing of these projects appeared in print during Newton's lifetime, it

Table 3. Summary by sections of Book 2 of *Principia*, 1st edition (1687) (pp. 236–400). The most notable variants are the emendations to Prop. 10 that Newton introduced as a consequence of Niklaus and Johann Bernoulli's criticisms (see [Hall, 1980] and Whiteside's commentary in [Newton *Papers*, vol. 8, 469–697]). Newton was extremely dissatisfied with the treatment of fluid resistance in Section 7, and this Section was completely rewritten: details in [Newton, 1972]; an English translation of the passages replaced or removed in the 2nd and 3rd eds. is in [Cohen and Smith, 2001, 299–313]; and an historical commentary on the relevance of such variants is [Smith, 2001].

	Title: On the Motion of Bodies, Second Book	
Sect.	**1st edition**	**Main variants**
1	Point-mass projectile resisted as the velocity.	
2	Point-mass projectile resisted as the velocity squared. Use of Taylor series. Lemma 2 on moments of products, etc.	Prop. 10 emended in 2nd ed. Scholium, Lemma 2: changes relative to dispute with Leibniz in 3rd ed.
3	Point-mass projectile resisted as $k_1 v + k_2 v^2$.	
4	Spiral trajectories.	
5	Hydrostatics, density of the atmosphere.	
6	Pendulum retarded oscillations.	
7	Resisted motion, solid of least resistance, efflux from a vessel.	Mostly rewritten in 2nd ed. General Scholium: emended and moved at the end of Section 6, values on air resistance changed.
8	Wave propagation, sound.	Scholium on sound changed in 2nd ed.
9	Refutation of vortex theory.	

is interesting to consider them since they reveal Newton's evaluations of his own mathematical methods for natural philosophy. For instance we know of projects of gathering all the mathematical Lemmas in a separate introductory section [Newton *Papers*, vol. 6, 600–609; Cohen, 1971, 171–172]. From David Gregory's retrospective memorandum of a visit he paid to Newton in May 1694 we learn about projects of expanding the geometrical Sections 4 and 5, Book 1, into a separate appendix on the 'Geometry of the Ancients', and of adding a treatise on the quadrature of curves as a second mathematical appendix in order to show the method whereby 'curves can be squared' [Newton *Papers*, vol. 6, 601; Cohen, 1971, 193–194, 345–349]. In general, after the publication of the *Principia*, Newton showed a concern for two problems. The first was to relate his mathematical methods for natural philosophy to the tradition of ancient geometry. The second was to relate them to the new analytical method of fluxions. Both problems became urgent and indeed an obsession for Newton after the inception of the priority quarrel with Leibniz. In fact, during

Table 4. Summary of preliminaries to Book 3 of *Principia*, 1st edition (1687) (pp. 401–404).

First column indicates Hypothesis number. These opening pages of Book 3 were considerably altered: both in the 2nd and 3rd eds. Newton reworked the wording and the astronomical data. The nine 'hypotheses' which open Book 3, 1st ed., are basic for the development of gravitation theory; they are, however, quite different in character. The first two hypotheses are methodological rules which justify the inductive generalization which leads to the establishment of universal gravitation: from 2nd ed. they became Rule 1 and 2. Newton added two further rules (Rule 3 in 2nd ed., Rule 4 in 3rd ed.) and grouped them together as 'Rules for the study of natural philosophy'. Hypothesis 3 (dropped in 2nd ed.) might have alchemical overtones. Hypothesis 4 states that 'the center of the system of the world is at rest'. Hypotheses 5–9 concern astronomical observations on the planetary system. It is likely that Newton was happy to get rid of so many occurrences of the term 'hypothesis', since in the General Scholium, at the end of Book 3 (2nd ed.), he stated his famous 'I do not feign hypothesis'.

Title: The System of the World, Third Book	
1st edition	**Main variants**
Page 401: Newton states that while Books 1–2 are mathematical, Book 3 is philosophical.	
1 (simplicity of nature).	→ Rule 1 in 2nd and 3rd ed.
2 (similar causes for similar effects).	→ Rule 2 in 2nd and 3rd ed.
3 (bodies' tansformations). Related to Cor. 2, Prop. 6, Book 3.	Dropped in 2nd ed.
	New rule 3 introduced in 2nd and 3rd edition (invariant observed qualities of bodies can be taken as qualities of all bodies).
	New rule 4 introduced in 3rd edition (props. gathered from phenomena by induction should be considered exactly or very nearly true).
4 (on center of World System).	→ Hypothesis 1, moved after Prop. 10, Book 3, in 2nd and 3rd eds.
5 (Jupiter's moons obey area and Kepler's 3rd law).	→ Phenomenon 1 in 2nd and 3rd eds.
	Phenomenon 2 introduced in 2nd and 3rd eds. (Saturn's moons obey area and Kepler's 3rd law).
6 (planets orbit around the Sun).	→ Phenomenon 3 in 2nd and 3rd eds.
7 (Sun's planets obey Kepler's 3rd law).	→ Phenomenon 4 in 2nd and 3rd eds.
8 (Sun's planets obey area law rel. to Sun).	→ Phenomenon 5 in 2nd and 3rd eds.
9 (Moon obeys area law rel. to Earth).	→ Phenomenon 6 in 2nd and 3rd eds.

Table 5. Summary of Propositions in Book 3 of *Principia*, 1st edition (1687) (pp. 405–511).

Book 3 contains 42 Propositions. First column indicates Proposition number. The most important variant is the addition of the concluding General Scholium in 2nd ed. In the General Scholium Newton deals with theological and methodological themes (see Snobelen [2001]). In the 2nd and 3rd eds Newton revised a great deal of numerical data: details can be found in [Newton, 1972].

	1st edition	**Main variants**
1	Kepler's area + 3rd laws prove Jupiter's moons $1/r^2$ attracted by Jupiter.	
2	Kepler's area + 3rd laws prove Sun's planets $1/r^2$ attracted by Sun.	
3	Kepler's area + 3rd laws prove Moon $1/r^2$ attracted by Earth.	
4	'Moon test' proves Moon gravitates towards Earth.	
5	Hypothesis 2 proves Sun's planets and Jupiter's moons gravitate towards Sun and Jupiter respectively.	It (viz. Rule 2) proves the same for Saturn's moons.
6	Equivalence of gravitational and inertial mass proved by pendulum experiments.	
7	Gravitation is universal and proportional to quantity of matter.	
8	Mutual gravitation of spherical masses.	Corollaries (weights on planets and planets' densities) rewritten in 2nd ed.
9	Gravity in the interior of planets.	
10	Planetary motions remain unaltered for a long time.	
11	Planetary system's center of mass is at rest.	
12	Motion of the Sun caused by planet's attractions.	
13	Planetary orbits are elliptical and area law valid.	
14	Aphelia and nodes of planets are at rest: Newton claims this is the best test for gravitation theory. Corollaries on stability of stars' system.	
15–16	Determination of diameters, aphelia, and eccentricities of planetary orbits.	
17	Moon's libration.	

Table 5. (*Continued*)

	1st edition	**Main variants**
18–20	Oblate shape of planets proved by balancing of polar and equatorial columns. Estimate of Earth's oblateness. Weight variation in function of latitude.	Major variants in 2nd and 3rd eds on measures of one degree on the meridian and length of seconds pendulum at different latitudes.
21	Precession of the equinoxes: theory.	
22–23	Inequalities of Moon, Jupiter's moons and Saturn.	
24	Tides caused by Moon's and Sun's attraction.	
25–35	Three Moon's inequalities: the 'variation', the motion of the nodes, the variation of the inclination.	Two props by Machin on Moon's nodes added in 3rd ed. after Prop. 33. Scholium, Prop. 35, altered in 2nd ed.: reference to calculation on Moon's apogee motion deleted, Horroxian model of Moon's motion briefly presented, general theory of Moon's inequalities described.
36–37	Tides: determination of Sun's and Moon's action on sea.	
38	Shape of the Moon.	
39	Precession of the equinoxes: numerical determination. Lemma 2: result obtained 'by the method of fluxions' [Newton, 1999, 884].	Demonstration altered in 2nd ed.
40–42	Determination of cometary orbits. Lemma 5: interpolation method.	New astronomical data added in 2nd and 3rd eds.
	page 511 Errata.	Moved after Chapter Index in 3rd ed.
		General Scholium added in 2nd ed.
		Alphabetical subject index added in 2nd ed.

the priority dispute, Newton found himself in a double trap. From one point of view, he wished to use the *Principia* as proof of his knowledge of calculus prior to the publication of G.W. Leibniz's *Nova methodus* (1684) (§4). This led him to state most of the propositions of the *Principia* had been found by means of the 'new analysis', even though they had been published in a different, 'synthetic', form. It was easy, he claimed, to revert the synthetic demonstrations into analytical form. On the other hand, Newton was convinced that his geometrical mathematical way was superior to Leibniz's reliance on algorithm.

It is notable that Newton planned to introduce some scholia, concerning the wisdom of the ancient Hebrews, Egyptians and Phoenicians, in Book 3. In his opinion, the ancient sages possessed a superior knowledge: they accepted atomism, the heliocentric system, and even had some awareness of universal gravitation [McGuire and Rattansi, 1966; Casini, 1984].

None of these projects materialized in the second and third editions of the *Principia*, which appeared in 1713 and 1726 respectively. The emendations and variations between these editions have been studied in detail by Koyré and Cohen [Newton, 1972]. Most notably, Newton revised some propositions (Corollary 1 to Proposition 13, Book 1; Proposition 10, Book 2; and large sections of the second Book). Hypotheses 1 and 2, on which the inductive generalizations of Book 3 are based, were expanded into four 'rules for the study of natural philosophy'. Newton also added an alternative geometrical representation of central force extending Proposition 6, Book 1, and the demonstrations on central force motion in Sections 2 and 3, Book 1. This new representation is based on the determination of curvature of the trajectory: the idea is that the total normal component of force at an arbitrary point is proportional to the square of speed and inversely proportional to the radius of curvature [Brackenridge, 1995]. As we know, this idea was not a novelty in Newton's approach to central forces and indeed appeared also in the first edition in Proposition 28, Book 3 [Nauenberg, 1994].

Notwithstanding these important variants, in broad outline the structure of the first edition remained unaltered. The number and order of the propositions, as well as the methods of proof, remained almost unchanged. The most striking differences between the first and later editions occur at the beginning and at the very end. The second and third editions contain, in fact, a Preface, signed by Roger Cotes, in which the objections of the Cartesians and the Leibnizians against the concept of gravitation are refuted. Cotes also considers the relationships between Newton's cosmology and religion: he maintains that the cosmology of the *Principia* avoids the dangers of Cartesian mechanicism and Leibnizian metaphysics. According to Cotes—who was speaking on Newton's behalf—, while the Continentals risk to reduce the world to a mechanism which can work without God's intervention, the cosmology of gravitation requires the wise and providential intervention of God. These themes are discussed also in a concluding General Scholium that Newton appended to the second edition. In these concluding pages he maintains that 'to treat of God from phenomena is certainly part of natural philosophy' [Newton, 1999, 943]. It is indeed anachronistic to narrowly read the *Principia* as a work on mathematical physics, since Newton's natural philosophy is deeply intertwined with theology, with alchemy, with his belief to be a rediscoverer of an ancient, forgotten, knowledge. Recent scholarship has established that Newton inserted many half-hidden hints to his heretical anti-trinitarism in the General Scholium [Snobelen, 2001]. The General Scholium contains Newton's famous pronouncement that on the cause of gravity he would not 'feign hypotheses'.

12 THE IMPACT OF THE *PRINCIPIA*

To evaluate the impact of the *Principia* is a momentous task. From one point of view, one could say that its influence has not expired yet, since Newtonian mechanics is still

adopted in many fields of science and since many problems faced by Newton—such as the three-body problem—are still open questions. However, such a statement is possible only by underevaluating the changes in the conceptions of mechanical principles, and in the sophistication of mathematical techniques, that occurred after Newton's death.

Basic concepts, such as the conservation of the angular momentum (of a system of particles) and the conservation of energy, were lacking from the *Principia*. Indeed, several mistakes that Newton did in studying rigid and fluid body dynamics (e.g., in his supposed refutation of the Cartesian vortex theory in Section 9, Book 2, and in his study of precession of equinoxes in Book 3) were due to his lack of understanding of these fundamental principles. Further, during the 18th century extremal principles, such as the principle of least action or that of virtual velocities, were proposed by Continental mathematicians as alternative foundations to mechanics. These alternatives do not belong to the Newtonian tradition, nonetheless they played a major role in the development of what came to be called 'Newtonian mechanics'. They rather belong to the conceptual framework of G.W. Leibniz (1646–1716). In fact, as early as 1687, Leibniz began reinterpreting the *Principia* in terms of his own cosmology (based on vortex theory), physics (where conservation of energy was adumbrated), matter theory (where infinite divisibility and elasticity, rather than atomism and hard impacts were basic), and mathematical language (carried on in terms of differential and integral calculus) [Truesdell, 1960; Bertoloni Meli, 1993].

The mathematical language in which Newton wrote his work became soon obsolete. After the works of mathematicians such as Pierre Varignon (1654–1722), Johann Bernoulli (1667–1748), Leonhard Euler (1707–1783), Alexis-Claude Clairaut (1713–1765), Jean le Rond d'Alembert (1717–1783), Joseph Louis Lagrange (1736–1813) and Pierre Simon Laplace (1749–1827), analytic mechanics was carried on in terms of ordinary and partial differential equations, and variational calculus, rather than on Newtonian geometric limit procedures [Blay, 1992]. This is not said to detract anything from the Newtonian achievement. Indeed, one can just think that before the *Principia* natural philosophers were able to mathematize just parabolic motions, the unresisted pendulum and uniform circular motions. In the *Principia* Newton was modelling topics such as the perturbations of planets, the motion of tides, and the motion of projectiles in resisting media.

The greatness of Newton's mathematical achievement was recognised by all his contemporaries, even by his worse enemies. On the other hand, the physics of gravitation met first with skepticism, especially on the Continent. How could a force operate at astronomic distances, in void and instantaneously? According to Continentals, such as Leibniz, Huygens and Johann Bernoulli, a mechanical explanation—analogous to Cartesian vortices—of the propagation of this action was needed. They thought that Newton had developed a beautiful mathematical theory about an unsound physical hypothesis. The success of the *Principia*, and its acceptance on the Continent, depended mostly upon two factors: the predictive success of its mathematical models, the presence of many fruitful mathematical open problems. Even people who did not endorse the physics of gravitation were struck by these two aspects. In the decades following Newton's death, thanks to gravitation theory, it was possible to predict planetary shapes, planets' deviations from purely Keplerian elliptical motions, and the return of comets. At the middle of the 18th century the Newtonian paradigm was accepted by most astronomers, and the worries about gravitation faded away,

even though the nature of an acting at a distance instantaneous interaction was not better understood. The fertility of the *Principia* as a repertoire of open mathematical problems should not be underestimated. In dealing with advanced topics, Newton had employed obscure, even flawed, mathematical methods. The more advanced parts of Newton's work became a source of inspiration for many 18th-century mathematicians. Attempts to improve on Newton's mathematical treatment of the three-body problem, of the determination of the solid of least resistance, or of the attraction of ellipsoids of revolution, carried on by men such as Euler, d'Alembert and Lagrange changed the scene of mathematics.

After their work the *Principia*'s mathematical methods ceased to be of interest for practising mathematicians. At the end of the 18th century the *Principia* was read only by a handful of erudite historians, even though knowledge of the first three sections of Book 1 and the first propositions of Book 3 was often required to university students [Warwick, 2003]. This is perhaps inevitable: a classic is a book that everybody would like to have already read. This is indeed a pity, since those who have read the great master know how beautiful is his geometrical language and how many powerful mathematical insights are still contained in his *Principia*.

BIBLIOGRAPHY

Bennett, J., Cooper, M., Hunter, M., Jardine, L. 2003. *London's Leonardo: The life and work of Robert Hooke*, Oxford: Oxford University Press.

Bertoloni Meli, D. 1993. *Equivalence and priority: Newton versus Leibniz*, Oxford: Clarendon Press.

Blay, M. 1992. *La naissance de la mécanique analytique: la science du mouvement au tournant des XVIIe et XVIIIe siècles*, Paris: Presses Universitaires de France.

Brackenridge, B.J. 1995. *The key to Newton's dynamics*, Berkeley, Los Angeles, London: California University Press.

Brackenridge, B.J. 2003. 'Newton's easy quadratures "omitted for the sake of brevity" ', *Archive for history of exact sciences*, *57*, 313–336.

Casini, P. 1984. 'Newton: the classical scholia', *History of science*, *22*, 1–58.

Chandrasekhar, S. 1995. *Newton's Principia for the common reader*, Oxford: Clarendon Press.

Cohen, I.B. 1971. *Introduction to Newton's Principia*, Cambridge: Cambridge University Press.

Cohen, I.B. 1999. 'A guide to Newton's *Principia*', in [Newton, 1999], 1–370.

Cohen, I.B. 2002. 'Newton's concepts of force and mass, with notes on the laws of motion', in I.B. Cohen and G.E. Smith (eds.), *The Cambridge companion to Newton*, Cambridge: Cambridge University Press, 57–84.

Cohen, I.B. and Smith, G.E. 2001. *Isaac Newton's natural philosophy*, Cambridge, MA, London: MIT Press.

De Gandt, F. 1995. *Force and geometry in Newton's Principia*, Princeton: Princeton University Press.

Densmore, D. 1995. *Newton's Principia, the central argument: translation, notes, and expanded proofs* (trans. and illust. by W.H. Donahue), Santa Fe, New Mexico: Green Lion Press.

DiSalle, R. 2002. 'Newton's philosophical analysis of space and time', in I.B. Cohen and G.E. Smith (eds.), *The Cambridge companion to Newton*, Cambridge: Cambridge University Press, 33–56.

Di Sieno, S. and Galuzzi, M. 1989. 'La quinta sezione del primo libro dei *Principia*: Newton e il "Problema di Pappo" ', *Archives internationales d'histoire des sciences*, *39*, 51–68.

Guicciardini, N. 1999. *Reading the Principia: the debate on Newton's mathematical methods for natural philosophy from 1687 to 1736*, Cambridge: Cambridge University Press.

Hall, A.R. 1980. *Philosophers at war: the quarrel between Newton and Leibniz*, Cambridge: Cambridge University Press.

Herivel, J.W. 1965. *The background to Newton's Principia: a study of Newton's dynamical researches in the years 1664–1684*, Oxford: Clarendon Press.

Inwood, S. 2002. *The man who knew too much: the strange and inventive life of Robert Hooke, 1635–1703*, London: Macmillan.

Kollerstrom, N. 1992. 'Thomas Simpson and "Newton's method of approximation": an enduring myth', *The British journal for the history of science*, *25*, 347–354.

Kollerstrom, N. 1999. 'The path of Halley's comet, and Newton's late apprehension of the law of gravity', *Annals of science*, *56*, 331–356.

McGuire, J.E. and Rattansi, M. 1966. 'Newton and the "Pipes of Pan" ', *Notes and records of the Royal Society of London*, *21*, 108–143.

Nauenberg, M. 1994. 'Newton's early computational method for dynamics', *Archive for history of exact sciences*, *46*, 221–252.

Nauenberg, M. 2001. 'Newton's perturbation methods for the three-body problem and their application to lunar motion', in J. Buchwald and I. B. Cohen (eds.), *Isaac Newton's natural philosophy*, Cambridge, MA, London: MIT Press, 189–224.

Nauenberg, M. 2004. 'Robert Hooke's seminal contribution to orbital dynamics', *Physics in perspective*, to appear.

Newton, I. *Correspondence*. *The correspondence of Isaac Newton*, 7 vols., H.W. Turnbull et alii (eds.), Cambridge: Cambridge University Press.

Newton, I. *Papers*. *The mathematical papers of Isaac Newton*, 8 vols., D.T. Whiteside et alii (eds.), Cambridge: Cambridge University Press.

Newton, I. 1972. *Philosophiae naturalis principia mathematica. The third edition (1726) with variant readings assembled and edited by Alexandre Koyré and I. Bernard Cohen, with the assistance of Anne Whitman*, Cambridge: Cambridge University Press.

Newton, I. 1999. *The Principia: mathematical principles of natural philosophy, a new translation by I. Bernard Cohen and Anne Whitman assisted by Julia Budenz, preceded by a guide to Newton's Principia by I. Bernard Cohen*, Berkeley, Los Angeles, London: University of California Press.

Pesic, P. 2001. 'The validity of Newton's lemma 28', *Historia mathematica*, *28*, 215–219.

Pourciau, B. 1991. 'On Newton's proof that inverse-square orbits must be conics', *Annals of science*, *48*, 159–172.

Pourciau, B. 2001. 'The integrability of ovals: Newton's lemma 28 and its counterexamples', *Archive for history of exact sciences*, *55*, 479–499.

Ruffner, J.A. 2000. 'On Newton's propositions on comets: steps in transition, 1681–84', *Archive for history of exact sciences*, *54*, 259–277.

Smith, G.E. 2001. 'The Newtonian style in Book II of the *Principia*', in [Cohen and Smith, 2001], 249–298.

Snobelen, S.D. 2001. ' "God of gods and Lords of lords": the theology of Isaac Newton's General Scholium to the *Principia*', *Osiris*, *16*, 169–208.

Truesdell, C. 1960. 'A program toward rediscovering the rational mechanics of the age of reason', *Archive for history of exact sciences*, *1*, 3–36.

Warwick, A. 2003. *Masters of theory: Cambridge and the rise of modern physics*, Chicago: University of Chicago Press.

Weinstock, R. 2000. 'Inverse-square orbits in Newton's *Principia* and twentieth-century commentary thereon', *Archive for history of exact sciences*, *55*, 137–162.

Whiteside, D.T. 1970. 'The mathematical principles underlying Newton's *Principia Mathematica*', *Journal for the history of astronomy*, *1*, 116–138.

Whiteside, D.T. 1991. 'The prehistory of the *Principia* from 1664 to 1686', *Notes and records of the Royal Society of London, 45*, 11–61.

Wilson, C. 2001. 'Newton on the Moon's variation and apsidal motion: the need for a newer "new analysis" ', in J. Buchwald and I.B. Cohen (eds.), *Isaac Newton's natural philosophy*, Cambridge, MA, London: MIT Press, 139–188.

CHAPTER 6

JAKOB BERNOULLI, *ARS CONJECTANDI* (1713)

Ivo Schneider

This book marks the unification of the calculus of games of chance and the realm of the probable by introducing the classical measure of probability. Justified by Bernoulli's law of large numbers, it contains a program to mathematize the realm of the probable, including what now is called the social domain.

First publication. *Ars conjectandi, Opus posthumum. Accedit tractatus de seriebus infinitis, et epistola Gallicè scripta de ludo pilae reticularis*, Basel: Thurnisii fratres, 1713. 306 + 35 pages.

Photoreprint. Brussels: *Editions Culture et Civilisation*, 1968.

New edition. In *Die Werke von Jakob Bernoulli*, vol. 3 (ed. and comm. B.L. van der Waerden and K. Kohli), Basel: Birkhäuser, 1975, 107–286.

Partial English translations. 1) With the Latin original, of chs. 1–3 of Part II in *The doctrine of permutations and combinations, being an essential and fundamental part of the doctrine of changes, as it is delivered by Mr. James Bernoulli,* [. . .] *and by the celebrated Dr. John Wallis* (trans. Francis Maseres), London: B. and J. White, 1795. 2) Of Part IV (trans. Bing Sung) as Harvard University, Department of Statistics, Technical Report No. 2 (1966); chs. 1–4 also available on the Web under http://cerebro.xu.edu/math/Sources/JamesBernoulli/ars_sung/ars_sung.html.

Full French translation. *L'art de conjecturer: suivi du Traité des séries infinies et de la Lettre sur le jeu de paume* (ed. Jean Peyroux), Paris: Blanchard, 1998.

Partial French translations. 1) Of Part I as *L'Art de conjecturer. . . avec des observations, éclaircissemens et additions. . . 1re partie* (trans. L.G.F. Vastel), Caen: [anonymous], 1801. [Includes also part of C. Huygens, *De la manière de raisonner dans les jeux de hazard*.] 2) Selection with Latin original in *Jacques Bernoulli et l'Ars conjectandi: documents pour l'étude de l'émergence d'une mathématisation de la stochastique* (trans. Norbert Meusnier), Mont Saint Aignan: Université de Rouen Haute Normandie, Institut de Recherche sur l'Enseignement des Mathématiques, 1987.

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

Nearly complete German translation. *Wahrscheinlichkeitsrechnung* (*Ars conjectandi*) (trans. R. Haussner), Leipzig: Engelmann, 1899 (*Ostwalds Klassiker der exakten Wissenschaften*, nos. 107 and 108). [Repr. Frankfurt/Main: Deutsch, 1999, 2002.]

Partial Italian translation. Of Part I as (trans. P. Dupont and Clara Silvia Roero) 'Il trattato "De ratiociniis in ludo aleae" di Christiaan Huygens con le "Annotationes" di Jacob Bernoulli ("Ars conjectandi", Parte I)', *Memorie della Accademia delle Scienze di Torino, (5) 8* (1984), 1–258.

Partial Russian translation. Of Part IV in *J. Bernoulli: On the law of large numbers* [in Russian] (trans. J.V. Uspensky, ed. A.A. Markov), Moscow: 1913. [Repr. (ed. with notes and commentaries by Yu.V. Prohorov), Moscow: Nauka, 1986.]

Partial Swedish translation. Of Part I (trans. Carl V. Ludvig Charlier) manuscript, University of Lund, 1919.

Related articles: de Moivre (§7), Bayes (§15), Laplace on probability (§24).

1 BACKGROUND AND STORY OF PUBLICATION

Born in 1755 in Basel as the son of a merchant, Jakob Bernoulli studied theology until he had received the licentiate in 1676. He left the subject in order to devote his time to astronomy and mathematics. Before he became professor of mathematics in Basel in 1687 he had travelled in France, the Netherlands, England, and Germany where he had met with mathematicians like Jan Hudde, John Wallis and Isaac Barrow. In these years he had become a Cartesian.

After G.W. Leibniz's publications concerning his form of the calculus in 1684 and 1686 (§4), Jakob and his younger brother Johann (1667–1748) began to contribute to the new field. Jakob cultivated the theory of infinite series on the basis of preliminary work done by Nikolaus Mercator, James Gregory, Isaac Newton, and Leibniz. He published five dissertations on series between 1689 and 1704. He considered series as the universal means to integrate arbitrary functions, to square and rectify curves.

One can only speculate why Jakob Bernoulli started in the mid 1680s to work on the calculus of games of chance and its extension to decision problems in everyday life. At this time only a few publications on the calculus of games of chance were available. Bernoulli's first source was the *De ratiociniis in ludo aleae* by Christiaan Huygens (1629–1695), which had come out in 1657 in Leiden as an appendix to Frans van Schooten's *Exercitationum mathematicarum libri quinque*. At this time the realm of games of chance and the realm of the probable which concerned opinion, evidence, and argument in practical problems of everyday life were still completely separated. Bernoulli bridged the gap between these two realms by combining Huygens's concept of expectation with a quantifiable concept of probability understood as degree of certainty in the *Ars conjectandi* (hereafter, 'AC'). According to his brother's and his own statements, he had begun work on the manuscript of AC in the 1690s, and had resumed work after some interruptions in the last two years of his life during which he corresponded with Leibniz about the main ideas of his art of conjecturing.

Leibniz himself had developed similar ideas independently of Jakob Bernoulli. He was interested in the creation of a doctrine or a logic of degrees of probabilities and hoped that, since his many obligations hindered him to do it himself, a mathematician like Jakob Bernoulli would indulge in the creation of this new theory of probability. When he had informed Leibniz in a letter from October 1703 about his intentions concerning his estimates of probabilities and especially his law of large numbers Leibniz reacted very critically. Leibniz's main criticisms were that the probability of contingent events, which he identified with dependence on infinitely many conditions, could not be determined by a finite number of observations and that the appearance of new circumstances could change the probability of an event. Bernoulli agreed that only a finite number of trials can be undertaken; but he differed from Leibniz in being convinced by the urn model that a reasonably great number of trials yielded estimates of the sought-after probabilities that were sufficient for all practical purposes.

Important sections of the AC were sketched out in Jakob Bernoulli's scientific diary, the 'Meditationes', from the mid 1680s onwards. When he died in 1705, the AC was not finished, especially lacking good examples for the applications of his 'art of conjecturing' to what he described as civil and moral affairs. Concerning the time that it would have needed to complete it, opinions differ from a few weeks to quite a few years, depending on assumptions about his own understanding of completeness. His heirs did not want his brother Johann, the leading mathematician in Europe at this time, to complete and edit the manuscript, fearing that Johann would exploit his brother's work. Only after Pierre Rémond de Montmort (1678–1719), himself a pioneer of the theory of probability, had sent an offer via Johann to print the manuscript at his own expense in 1710, and after some admonitions that the *Ars conjectandi* soon would become obsolete if not published, Jakob's son, a painter, agreed to have the unaltered manuscript printed. It appeared in August 1713 together with a tract about infinite series and a letter in French on the 'Jeu de paume', a predecessor of tennis.

A short preface was contributed by Nikolaus Bernoulli (1687–1759), Jakob's nephew. He had read the manuscript when his uncle was still alive, and had made considerable use of it in his thesis of 1709 and in his correspondence with Montmort. He was asked twice to complete and edit the manuscript. The first time he excused himself by his absence when he travelled in 1712 to Holland, England and France. After his return Nikolaus Bernoulli declared himself as too inexperienced to do the job and in his preface he asked Montmort, the anonymous author of the *Essay sur les jeux de hazard*, and Abraham de Moivre (1667–1754) to complete his uncle's work.

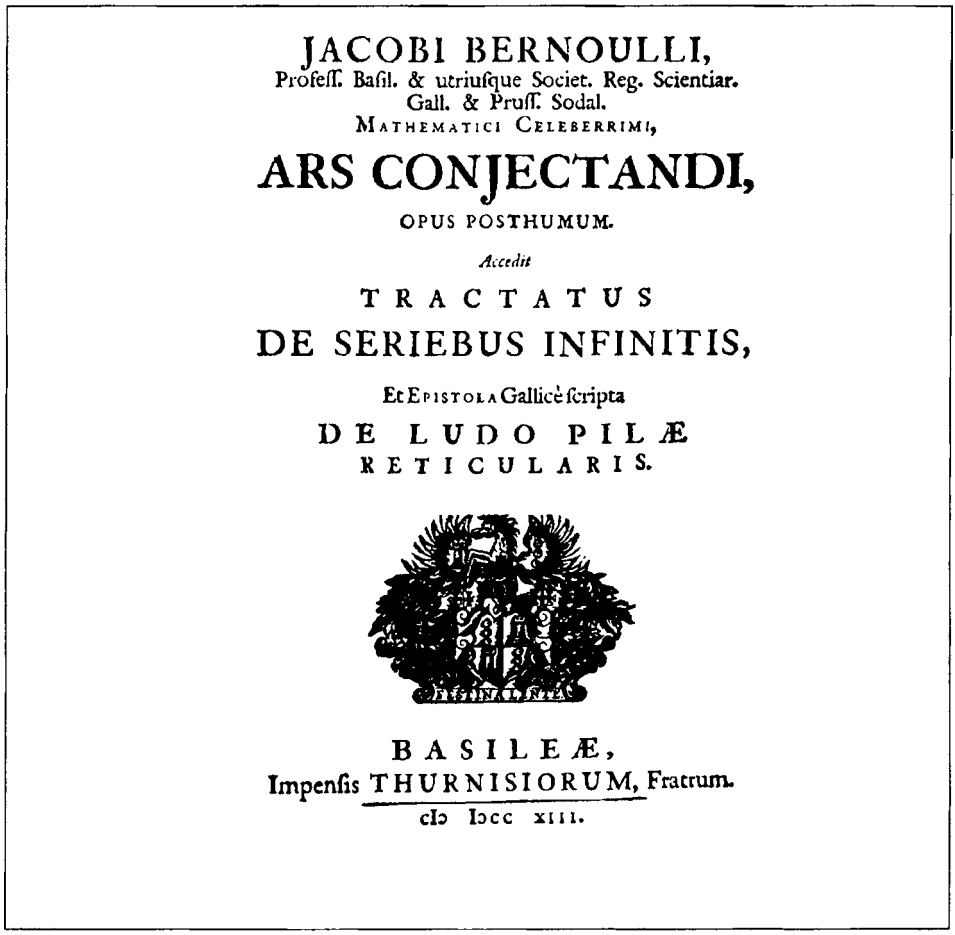
2 CONTENT AND STRUCTURE OF THE AC

After an attractive title page (Figure 1), the *Ars conjectandi* consists of four Parts (Table 1). The first Part (pp. 1–71) is based on Huygens's *De ratiociniis in ludo aleae aleae*. The second Part (pp. 72–137) deals with the theory of permutations and combinations. The third (pp. 138–209) contains 24 problems concerning games of chance. The fourth and last (pp. 210–239) is devoted to the application of the art of conjecturing to 'civil, social, and economical affairs'. It is followed by a tract on infinite series (pp. 241–306) which is

Table 1. Contents by chapters of Bernoulli's book.

Part/Ch.	**P.**	**Topics**
Preface by Nikolaus Bernoulli		
1st Part	1	Huygens's preface to his *De ratiociniis in ludo aleae*.
	3	Huygens's propositions I to III with Bernoulli's comments.
	11	The problem of points in Huygens's Propositions IV–IX.
	20	Huygens's dicing problems in Propositions X–XII.
	38	Bernoulli's generalization leading to the binomial distribution.
	45	Huygens's Propositions XIII and XIV.
	49	Bernoulli's solution of Huygens's problems I–V.
2nd Part	72	*Introduction to the theory of permutations and combinations.*
Ch. 1	74	Permutations.
Ch. 2	82	Combinations of all classes.
Ch. 3	86	Combinations of a particular class; figurate numbers; the general formula for sums of powers of integers (p. 97).
Ch. 4	99	Properties of the binomial coefficients and the treatment of the problem of points for two players with equal chances for a win.
Ch. 5	112	Combinations with repetitions.
Ch. 6	118	Combinations with restricted repetitions.
Ch. 7	124	Variations.
Ch. 8	127	Variations with repetitions.
Ch. 9	132	Variations with restricted repetitions.
3rd Part	138	*Application of the theory of permutations and combinations to 24 problems dealing with games of chance, including Bassette* (pp. 191–199).
4th Part	210	*The application of the theory of permutations and combinations in the social, political and economical domain.*
Ch. 1	210	About certitude, probability, necessity and contingent events.
Ch. 2	213	About science and conjecture.
Ch. 3	217	The estimation of the weight of different arguments.
Ch. 4	223	The two ways to determine probabilities, especially by often repeated trials.
Ch. 5	228	Proof of Bernoulli's law of large numbers.
App. 1	241	Tract concerning the summation of infinite series and their application to quadratures and rectifications. [End 306.]
App. 2	(1)	'Lettre à un Amy, sur les Parties du Jeu de Paume'. [End (35).]
App. 3	–	List of Errata.

however without any connection to games of chance or probability theory and has been treated in histories of the infinitesimal calculus in the 17th century. A 35-page appendix with separate pagination deals with the jeu de paume. On the context of the book see [Todhunter, 1865, ch. 7; Hacking, 1975, chs. 16–17; Stigler, 1986, ch. 2; and Hald, 1990, chs. 15–16].



JACOBI BERNOULLI,
Profeſſ. Baſil. & utriuſque Societ. Reg. Scientiar.
Gall. & Pruſſ. Sodal.
MATHEMATICI CELEBERRIMI,

ARS CONJECTANDI,

OPUS POSTHUMUM.

Accedit

TRACTATUS
DE SERIEBUS INFINITIS,

Et EPISTOLA Gallicè ſcripta
DE LUDO PILÆ
RETICULARIS.

BASILEÆ,
Impenſis THURNISIORUM, Fratrum.
cIↃ IↃcc xiii.

Figure 1. Title page of Bernoulli's book.

3 HUYGENS'S *DE RATIONCINIIS IN LUDO ALEAE* WITH BERNOULLI'S ANNOTATIONS

In the first Part Jakob Bernoulli complemented his reprint of Huygens's tract by extensive annotations which contained important modifications and generalisations. Bernoulli's additions to Huygens's tract are about four times as long as the original text.

The central concept in Huygens's tract is expectation. The expectation of a player *A* engaged in a game of chance in a certain situation is identified by Huygens with his share of the stakes if the game is not played or not continued in a 'just' game. For the determination of expectation Huygens had given three propositions which constitute the 'theory' of his calculus of games of chance. Huygens's central proposition III maintains:

> If the number of cases I have for gaining a is p, and if the number of cases I have for gaining b is q, then assuming that all cases can happen equally easily, my expectation is worth $(pa + qb)/(p + q)$.

Bernoulli not only gives a new proof for this proposition but also generalizes it in several ways. In his Corollary 3 he secures the possibility to determine the expectation E if there are p_i cases for gaining a_i, $i = 1, \ldots, n$, beginning with $i = 2$ and proceeding to $i = 3$ and so on up to $i = n$ as $E = \frac{\sum_{i=1}^{n} p_i a_i}{\sum_{i=1}^{n} p_i}$. Bernoulli, however, did not use the same notation, as one can see from Corollary 3, which begins with the statement (p. 9):

> If I have p cases for gaining a, q cases for gaining b, and r cases for gaining c this is worth for me as much as if I would have $p + q$ cases for gaining $(pa + qb)/(p + q)$ and r cases for gaining c.

Applying Huygens's proposition III leads to the result $E = (pa + qb + rc)/(p + q + r)$, from which one can proceed to add s cases for gaining d, and so on.

Differently from Huygens, Bernoulli admitted that the 'gains' $a, b, c, \ldots$ can assume non-positive values, especially zero. So if in Huygens proposition III b is zero than $E = \frac{p}{p+q}a$, or the value of Huygens's expectation is the product of what Bernoulli will call in Part IV the measure of probability of an event leading to the gain a and a. If $a = 1$, as is assumed in the following by Bernoulli, Huygens's expectation and the corresponding measure of probability become numerically but still not conceptionally equal.

Huygens's propositions IV to VII treat the problem of points, also called the problem of the division of stakes, for two players; propositions VIII and IX treat three and more players. Bernoulli returns to these problems in Part II of the AC. In his annotations to Huygens's proposition IV he generalised Huygens's concept of expectation which, like the shares, is measured in terms of money. Bernoulli wanted to extend the meaning of share e.g. to include 'some kind of prize, laurels, victory, social status of a person or thing, public office, some kind of work, life or death'. Here he remarks that the expectations of two players have to add up to the total stakes if the corresponding events leading to win or loss are disjoint and complementary. This is not to be confused with the addition rule for probabilities. Bernoulli constructs an example where the respective expectations of two players do not add up to the total stakes.

This is the only instance in the annotations and commentaries to Huygens tract where Bernoulli uses the word 'probabilitas', or probability as understood in everyday life. Later in Part IV of the AC Bernoulli replaced Huygens's main concept, expectation, by the concept of probability for which he introduced the classical measure of favourable to all possible cases.

The remaining propositions X to XIV of Huygens's tract deal with dicing problems of the kind: What are the odds to throw a given number of points with two or three dice? or: With how many throws of a die can one undertake it to throw a six or a double six? In proposition XII Huygens solved the problem 'To find how many dice should one take to throw two sixes at the first throw'. In his extensive annotations Bernoulli generalizes the problem in different ways. In the first he does not restrict the number of dice to two or three. He solves the problem to determine the odds to throw a given number of points with n dice by an algorithm which he works out up to $n = 6$.

The meaning of Huygens's result of proposition X that the expectation of a player who contends to throw a six with four throws of a die is greater than that of his adversary is explained by Bernoulli in a way which relates to the law of large numbers proved in Part IV of the AC. Bernoulli's generalization of Huygens's proposition XI and XII leads to the problem of finding the expectation or, since the stakes are normalized to the amount 1, the probability, as introduced in Part IV of the AC, of a player who contends to achieve in a series of n independent trials at least m successes if the chances for success and failure are as $b : c$ and $b + c = a$ or the probability of success is b/a and that of failure c/a. Bernoulli finds inductively that this expectation or probability is

$$\sum_{\nu=0}^{n-m} \binom{n}{m+\nu} \left(\frac{b}{a}\right)^{m+\nu} \left(\frac{c}{a}\right)^{n-m-\nu}. \tag{1}$$

For this he considers the expectation $B(m, n)$ of the opponent who contends that there will be no more than $m - 1$ successes in n independent trials. He uses the reduction formula

$$B(m, n) = \frac{b \cdot B(m-1, n-1) + c \cdot B(m, n-1)}{a}, \tag{2}$$

where

$$B(0, n) = 0, \;\; B(1, n) = (c/a)^n, \text{ and } B(n, n) = 1 - (b/a)^n \quad \text{for all } n \geqslant 1. \tag{3}$$

He calculates $B(\nu, \mu)$, which he tabulates for $\nu \leqslant 4$ and $\mu \leqslant 6$ and extrapolates by incomplete induction

$$B(m, n) = \sum_{\mu=0}^{m-1} \binom{n}{\mu} \left(\frac{b}{a}\right)^{\mu} \left(\frac{c}{a}\right)^{n-\mu}. \tag{4}$$

He indicates how the same result can be achieved with the help of the theory of combinations which was not used by Huygens and which he is going to develop in the second Part.

Bernoulli's procedure presupposes the equivalent of the multiplication rule for independent events which he formulates most explicitly on p. 44. Following the multiplication rule he determines the probability of exactly m successes in n independent trials if the probability of success is b/a and that of failure c/a, and if the order of successes and failures does not matter as

$$\binom{n}{m} \frac{b^m c^{n-m}}{a^n} = \binom{n}{n-m} \frac{b^m c^{n-m}}{a^n}. \tag{5}$$

This is the binomial or, as it was also called later, the 'Bernoulli' distribution.

In proposition XIV Huygens wanted to know the odds of two players A and B the first of which wins if he throws seven points with two dice and the second if he throws six points with two dice and if the first throw is conceded to B. If the expectation of A for the stakes a is x as often as it is B's turn and y as often as it is his own turn, we get the two

equations

$$\frac{31}{36}y = x \text{ or } y = \frac{36}{31}x \text{ and } \frac{6a + 30x}{36} = y \text{ with the solution } x = \frac{31}{61}a. \tag{6}$$

Accordingly the odds for A and B are as $31 : 30$.

Bernoulli comments that this is the first time Huygens is forced to use analysis in order to find the solution. He maintains, however, that he can avoid analysis in the following way. He introduces infinitely many players in succession each having one attempt with probability $\frac{b}{a}$ for success and $\frac{c}{a} = \frac{a-b}{a}$ for failure for all odd-numbered players and with probability $\frac{e}{a}$ for success and $\frac{f}{a} = \frac{a-e}{a}$ for failure for all even-numbered players. The expectations of the first, second, third, ... players are accordingly

$$\frac{b}{a},\ \frac{ce}{a^2},\ \frac{bcf}{a^3},\ \frac{c^2ef}{a^4},\ \frac{bc^2f^2}{a^5},\ \frac{c^3ef^2}{a^6},\ \frac{bc^3f^3}{a^7},\ \frac{c^4ef^3}{a^8},\ \ldots. \tag{7}$$

If one replaces all odd-numbered players by the single player B and all even-numbered players by the player A, then by the addition rule, the expectation of B is

$$E(B) = \sum_{\nu=0}^{\infty} \frac{b}{a}\left(\frac{cf}{a^2}\right)^{\nu} = \frac{ab}{a^2 - cf} \quad \text{and} \quad E(A) = \sum_{\nu=0}^{\infty} \frac{ce}{a^2}\left(\frac{cf}{a^2}\right)^{\nu} = \frac{ce}{a^2 - cf}. \tag{8}$$

The odds for A and B are therefore as $ce : ab$.

Proposition XIV was the last in Huygens's tract. He had added five problems with the results for problems 1, 3, and 5 but without indicating how he had achieved these results; problems 2 and 4 are without either result or any hint how to solve them. Huygens liked to leave to his readers the solution of these five problems, which were far more complicated than those he had solved in his tract following the example of the mathematical practitioners and reckoning masters of the preceding and his own century who challenged one another with the most difficult problems. These problems at the end of Huygens's tract became kind of a training program for all interested in the calculus of games of chance in the generation after Huygens like Baruch Spinoza, de Moivre, Montmort, Nicolaas Struyck, and of course Jakob Bernoulli.

Huygens's first problem can be solved with the same methods applied for the solution of proposition XIV. However, Bernoulli shows that his method with infinitely many players and infinite series is not restricted as with Huygens to the periodic case where the same probabilities of success and failure occur in a certain order. In Huygens's second problem one has to find the odds of three players A, B, and C who draw blindly in the order $ABCABC\ldots$ from a set of 12 chips, 8 black and 4 white, until the first wins by drawing a white chip. Bernoulli first emphasizes that the problem as stated by Huygens allows for three different interpretations, each of which is solved following Huygens's method and with Bernoulli's own method of infinite series. Bernoulli solves Huygens's third problem first by Huygens's method of recursion and then by combinatorics. The fourth problem is solved in the third Part of the AC.

Huygens's fifth problem is a special case of the gambler's ruin problem. It asks for the chances of two players A and B, the first having n and the second m counters and their

respective chances for winning in every single trial being as $p : q$, $q = 1 - p$, to be ruined that is to say to lose all respective counters if the loser of a trial has to give one counter to the winner. Bernoulli solves this problem for the case $m = n = 12$ by applying Huygens's method with the result that the chances of A and B are as $p^{12} : q^{12}$. In addition, he states without proof that the chances of A and B in the general case of n respectively m counters are as

$$\left(p^n q^m - p^{m+n}\right) : \left(q^{m+n} - p^n q^m\right). \tag{9}$$

4 COMBINATORICS AS THE MAIN TOOL OF THE ART OF CONJECTURING

In the second Part Bernoulli deals with combinatorial analysis, based on contributions of van Schooten, Leibniz, Wallis, and Jean Prestet. Later when dealing with figurate numbers he adds the names of Johannes Faulhaber, Johannes Remmelin, and Nicolaus Mercator. In connection with the general formula for sums of powers of integers he mentions Ismaël Boulliau, who took some hundreds of pages in order to find these sums up to the exponent 6. To these we can add a series of authors who dealt with permutations of the letters which constitute single words or whole phrases under condition that the new words or phrases make sense or if it are verses that the metric is preserved.

The second Part consists of nine chapters dealing with permutations, the number of combinations of all classes, the number of combinations of a particular class, figurate numbers and their properties (especially the multiplicative property), sums of powers of integers, the hypergeometric distribution, the problem of points for two players with equal chances to win a single game, combinations with repetitions and with restricted repetitions, and variations with repetitions and with restricted repetitions.

Evidently Bernoulli did not know Blaise Pascal's *Triangle arithmétique*, published posthumouly in 1665, though Leibniz had alluded to it in his last letter to him in 1705. Not only does Bernoulli not mention Pascal in the list of authors that he had consulted concerning combinatorial analysis except for Pascal's letter to Fermat of 24 July 1654; it would also be difficult to explain why he repeated results already published by Pascal in the *Triangle arithmétique,* such as the multiplicative property for binomial coefficients for which Bernoulli claims the first proof for himself. His arrangement differs completely from that of Pascal, whose proof for the multiplicative property of the binomial coefficients has been judged to be clearer than Bernoulli's [Edwards, 1987, 134]. It is fair to add that in the AC, which Bernoulli left as an unpublished manuscript, he was much more honest concerning the achievements of his predecessors than Pascal in the *Triangle arithmétique*. It is also true that Bernoulli was concerned with combinatorial analysis in the AC first of all because it constituted for him a most useful and indispensable universal instrument for dealing numerically with conjectures, since 'every conjecture is founded upon combinations of the effective causes' (p. 73).

Chapter I deals with permutations stating the rule that there are $n!$ different arrangements or permutations of n different things and $n!/(a!b!c!\cdots)$ permutations of a things of one kind, b of another, c of another, and so on with $a + b + c + \cdots = n$. In chapter II he gives the rule for finding the number of combinations of n different things taken one or more at a time. Chapter III offers a form of the arithmetical triangle in which the number

in the nth row and rth column represents the number of combinations ${}^{n-1}C_{r-1}$ of order $r-1$ formed from $n-1$ different things. Bernoulli identifies these numbers with the figurate numbers and their properties equivalent, for example, to ${}^{n}C_r = \sum_{i=1}^{n-r+1} {}^{n-i}C_{r-1}$. He states also that the numbers in the $(n+1)$th row beginning with the first column constitute the coefficients of the binomial $(1+1)^n = \sum_{i=0}^{n} \binom{n}{i}$ and proves that

$$ {}^{n}C_r = \frac{\prod_{\nu=0}^{r-1}(n-\nu)}{\prod_{\nu=0}^{r-1}(r-\nu)}. \tag{10} $$

Most of these results can be found in the works of other authors even before Pascal. As [Schneider, 1993, chs. 5 and 7] has shown, all these results can be found in the published work of Johannes Faulhaber (1580–1635). Over and above that one can find in Faulhaber's works the sums of powers of integers up to exponent 17 in the following form:

$$ \sum_{\nu=1}^{n} \nu^{2s+1} = \left(\sum_{\nu=1}^{n} \nu \right)^2 \cdot \left[\sum_{i=1}^{s} a_i \left(\sum_{\nu=1}^{n} \nu \right)^{s-i} \right] \quad \text{for odd exponents} \tag{11} $$

and

$$ \sum_{\nu=1}^{n} \nu^{2s} = \sum_{\nu=1}^{n} \nu^2 \cdot \left[\sum_{i=1}^{s} b_i \left(\sum_{\nu=1}^{n} \nu \right)^{s-i} \right] \quad \text{for even exponents,} \tag{12} $$

with appropriate algorithms for the determination of the coefficients a_i and b_i.

Apart from the fact that this kind of representation of the sums of powers of integers in form of polynomials in $\sum_{\nu=1}^{n} \nu$ is all but trivial and that Faulhaber had found corresponding formulas for higher sums of powers of integers, it is easy to transform Faulhaber's polynomials in $\sum_{\nu=1}^{n} \nu$ into polynomials in n and deduce from them general properties of the coefficients. However, Faulhaber had not done this.

According to the stocks of the University Library in Basel, Jakob Bernoulli seems not to have had access to those publications of Faulhaber that contained the formulas for the sums of powers of integers. But he used his findings concerning combinatorial analysis in order to develop a general formula for the sums of powers of integers in the following way: With ${}^{n}C_r = \sum_{i=1}^{n-r+1} {}^{n-i}C_{r-1}$ and the product property of the ${}^{n}C_r$ he can find successively the sums $\sum_{\nu=1}^{n} \nu^c$ beginning with $c = 1$ according to

$$ {}^{n}C_2 = \frac{n \cdot (n-1)}{2} = \sum_{i=1}^{n-1} {}^{n-i}C_1 = \sum_{i=1}^{n-1} (n-i) = \sum_{i=1}^{n-1} i \quad \text{or} \quad \sum_{i=1}^{n} i = \binom{n+1}{2}. \tag{13} $$

In this way he determines the sums of powers of integers up to $c = 10$ as polynomials in n which he displays in a table and which he takes as an induction basis for the general

formula

$$\sum_{\nu=1}^{n}\nu^c=\frac{1}{c+1}n^{c+1}+\frac{1}{2}n^c+\frac{1}{2}\binom{c}{1}An^{c-1}+\frac{1}{4}\binom{c}{3}Bn^{c-3}+\frac{1}{6}\binom{c}{5}Cn^{c-5}$$
$$+\frac{1}{8}\binom{c}{7}Dn^{c-7}+\cdots \tag{14}$$

where $A, B, C, D, \ldots$ are the coefficients of n in $\sum_{\nu=1}^{n}\nu^2$, $\sum_{\nu=1}^{n}\nu^4$, $\sum_{\nu=1}^{n}\nu^6$, $\sum_{\nu=1}^{n}\nu^8, \ldots$ with the values $A=1/6$, $B=-1/30$, $C=1/42$, $D=-1/30, \ldots$. He observes that the sum of the coefficients in each polynomial for $\sum_{\nu=1}^{n}\nu^c$ for all $c\geqslant 1$ must be 1, which is evident by putting $n=1$ in the respective formulas. This allows him to calculate $A, B, C, D, \ldots$ successively. These constants were later called 'Bernoulli numbers' by de Moivre and Euler (compare §7 and §13). Bernoulli's formula for $\sum_{\nu=1}^{n}\nu^c$ played an important role not only for the demonstration of de Moivre's form of the central limit theorem but also for analysis in general.

At the end of chapter III Bernoulli shows for $r=3$ how to determine the n th term and the sum of the first n terms of an arithmetical series of order r. Even this case of the so called Newton–Gregory formula can be found already in Faulhaber's *Academia Algebrae* (1632).

Chapter IV contains further properties of the binomial coefficients such as

$${}^{n}C_r={}^{n-1}C_r+{}^{n-1}C_{r-1} \text{ being equivalent to } \binom{n}{r}=\binom{n-1}{r}+\binom{n-1}{r-1}; \tag{15}$$

this is the combinatorial treatment of the problem of points for two players A and B with equal chances for a single win the first lacking n and the second m wins. The chances of A to win the whole game are given as $\left(\frac{1}{2}\right)^{n+m-1}\sum_{\nu=n}^{n+m-1}\binom{n+m-1}{\nu}$. Pascal's equivalent solution of the same problem can be reconstructed from his letters to Fermat in 1654 and is certainly contained in the *Triangle arithmétique*.

In chapter V Bernoulli deals with ${}^{s}_{(r)}C_r$, the combinations of r things with repetitions from $s\geqslant r$ different things, for which he gets in a table of combinations with repetition another arrangement of the arithmetical triangle for terms of which satisfy

$${}^{s}_{(r)}C_r=\sum_{i=0}^{r-1}{}^{s-1}_{(r)}C_{r-i}+1=\binom{s+r-1}{r}. \tag{16}$$

The remaining four chapters deal with combinations with restricted repetitions, variations without and with repetitions and with restricted repetitions. The most interesting detail is contained in chapter VIII, where he states that the number of variations of m things with repetitions out of n different things is given by the coefficients of the multinomial expansion

$$(a_1+\cdots+a_n)^m=\sum_{\substack{0\leqslant r_i\leqslant m\\ r_1+\cdots+r_n=m}}\frac{m!}{r_1!\cdots r_n!}a_1^{r_1}\cdots a_n^{r_n}. \tag{17}$$

Except for Bernoulli's general formula for powers of integers, even those results that he considered as new can be found in the works of other authors, especially in Pascal's *Triangle arithmétique*. Nevertheless, his exposition of combinatorics in the AC became the most popular text in the 18th century [Hald, 1990, 229]. This might be due to its well-structured presentation and its interweaving with the calculus of games of chance.

5 NEW PROBLEMS AND EXERCISES IN THE THIRD PART

In the third Part Bernoulli gives 24 problems concerning the determination of the modified Huygenian concept of expectation in various games. Here he uses extensively conditional expectations without, however, distinguishing them from unconditional expectations. All the games are games of chance with dice and cards including games en vogue at the French court of the time like Cinque et neuf, Trijaques, or Basette. He solves these problems mainly by combinatorial methods, as introduced in Part II, and by recursion.

Typical for the use of conditional expectations are games constructed as a two stage experiment, where the outcome of the first stage determines the conditions for the experiment at the second stage. The first form of problem 14 is representative for his procedure: Two players A and B agree that A throws a die and in a second stage the same die a number of times which corresponds to the number of points he achieved with the first throw. They stake together the amount 1. A will win 1, $1/2$, or 0 if the number of points he made at the second stage is > 12, $= 12$, or < 12. Bernoulli than determines the six conditional expectations of A $E(A|x)$, $x = 1, \ldots, 6$, depending on the outcome of the first throw. These are found by counting the cases leading to a number of points which is 12, $= 12$, or < 12. So $E(A|1) = 0$ because with one throw of a die one can achieve not more than 6 points that is less than 12 points; and because there is only one case among 36 in which 12 points can be achieved $E(A|2) = \frac{1}{72}$. From the 216 cases one has with three dice, 25 lead to 12 points and 56 to more than 12 points; accordingly,

$$E(A|3) = \frac{25 \cdot \frac{1}{2} + 56 \cdot 1}{216} = \frac{137}{432}, \tag{18}$$

etc. The unconditional expectation of A is than $E(A) = \sum_{x=1}^{6} E(A|x)/6$ because there is just one case for every conditional expectation.

From a methodological point of view these problems do not offer anything especially new. However, they represent the standard of the problems treated at about the same time by, for example, Joseph Sauveur, Montmort and de Moivre; some of these problems or modifications of them reappear in the works of these authors.

6 THE TRANSITION TO A CALCULUS OF PROBABILITIES IN THE FOURTH PART

This is the most interesting and original Part; but it is the one that Bernoulli was not able to complete. In the first three of its five chapters it deals with the new central concept of the art of conjecturing, probability, its relation to certainty, necessity and chance, and ways of

estimating and measuring probability. Jakob Bernoulli states on the one hand that, at least for God, chance and with it objective probabilities do not exist in a world the functioning of which in the past, present, and future is completely known to him down to its smallest entities. Through a more precise knowledge of the parameters affecting the motion of a die, for instance, it would be possible even for men to specify in advance the result of the throw. Chance, in his view and later in the view of Laplace, was reduced to a subjective lack of information. Thus, depending on the state of their information, an event may be described by one person as chance, but by another as necessary. The entire realm of events which are described in daily life as uncertain or contingent in their outcome is such, he claims, merely because of incomplete information: nevertheless, these too are covered by his concept of probability which he introduces as follows (p. 211):

> For probability is a degree of certainty and differs from it as a part differs from the whole. If, for example, the whole and absolute certainty—which we designate by the letter a or by unity—is supposed to consist of five probabilities or parts, three of which stand for the existence or future existence of some event, the remaining against, this event is said to have $\frac{3}{5}a$ or $\frac{3}{5}$ certainty.

However, the only way to check if the guessed, estimated or calculated probabilities of events are reliable in practice is to make sufficiently many observations and calculate the relative frequencies of their outcome. So the denial of objective chance and probabilities does not prevent Bernoulli's concept of probability, defined as degree of certainty, from showing at the same time subjective or epistemic and frequentist or aleatory aspects.

For the mathematical part of the *Ars conjectandi* this ambiguity in Bernoulli's concept of probability does not matter, because the numerical value of a probability is a real—in Bernoulli's case a rational—number between zero and one, no matter how it is conceived. For practical applications he introduces the concept of moral certainty of events, 'whose probability nearly equals the whole of certainty'.

In chapter 3, 'About various kinds of arguments and how their weights are estimated for quantifying the probabilities of things' the realm of non-additive probabilities is touched [Shafer, 1978, 323–341]. In chapter 4 Bernoulli distinguishes two ways of determining, exactly or approximately, the classical measure of probability. The first presupposes equipossibility of the outcomes of certain elementary events like drawing either one of n balls numbered from 1 to n out of an urn. So the probability of drawing a ball of a certain colour out of an urn filled with balls of different colours is determined *a priori* by the ratio of the number of balls of this special colour to the number of all balls in the urn. For the determination of the probability of an event like a certain person's dying within the next ten years a reduction to numbers of equipossible cases which are favourable or unfavourable for the event is impossible. But according to Bernoulli we can inductively, by experiments, or *a posteriori* in his sense, get as close as we desire to the true measure of such a probability. The possibility of estimating the unknown probability of such an event by the relative frequency of the outcome of this event in a series of supposedly independent trials is secured, according to Jakob Bernoulli, by his *theorema aureum* ('golden theorem'), which was called later by S.D. Poisson 'Bernoulli's law of large numbers'. The proof of this theorem is contained in chapter 5. In it he had shown that the relative frequency h_{nt} of an event with probability $p = r/t$, $t = r + s$, in nt independent trials converges in probability to p.

More precisely, he had shown that, for any given small positive number $\varepsilon = 1/t$ and any given large natural number c, for sufficiently large n the inequality

$$\frac{\Pr\{|h_{nt} - p| \leqslant 1/t\}}{\Pr\{|h_{nt} - p| > 1/t\}} > c \tag{19}$$

holds, which is equivalent to

$$\Pr\left\{|h_{nt} - p| \leqslant \frac{1}{t}\right\} > \frac{c}{c+1} \quad \text{or} \quad 1 > \Pr\left\{|h_{nt} - p| \leqslant \frac{1}{t}\right\} > 1 - \frac{1}{c+1}, \tag{20}$$

which again is what is now called 'Bernoulli's weak law of large numbers'.

For the proof Bernoulli considered the binomial

$$\left(\frac{r}{t} + \frac{s}{t}\right)^{nt} = t^{-nt} \cdot (r+s)^{nt} = \sum_{i=-nr}^{ns} T_i, \quad T_i = \binom{nt}{nr+i} r^{ns-i} s^{nr+i} t^{-nt}. \tag{21}$$

The probability that the relative frequency h_{nt} does not deviate more than $1/t$ from $p = \frac{r}{t} = \frac{nr}{nt}$ is than $\Pr\{|h_{nt} - p| \leqslant 1/t\} = \sum_{i=-n}^{n} T_i$.

Bernoulli shows in four lemmas that

1. $T_0 = \max\{T_i\}$; (22)
2. $T_0 > T_{-1} > T_{-2} > \cdots > T_{-nr}$ and $T_0 > T_1 > T_2 > \cdots > T_{ns}$; (23)
3. $\dfrac{T_0}{T_{-1}} < \dfrac{T_{-1}}{T_{-2}} < \dfrac{T_{-2}}{T_{-3}} < \cdots < \dfrac{T_{-nr+1}}{T_{-nr}}$ and $\dfrac{T_0}{T_1} < \dfrac{T_1}{T_2} < \dfrac{T_2}{T_3} < \cdots < \dfrac{T_{ns-1}}{T_{ns}}$; (24)
4. $\dfrac{T_0}{T_n} < \dfrac{T_i}{T_{n+i}}$ and $\dfrac{T_0}{T_{-n}} < \dfrac{T_{-i}}{T_{-n-i}}$ for $i > 0$. (25)

It remains to prove that

$$\sum_{i=-n}^{n} T_i \geqslant c \cdot \left[\sum_{i=-nr}^{-n-1} T_i + \sum_{i=n+1}^{ns} T_i\right]. \tag{26}$$

According to lemmas 2 and 4 he gets

$$\frac{\sum_{i=1}^{n} T_i}{\sum_{i=n+1}^{ns} T_i} \geqslant \frac{\sum_{i=1}^{n} T_i}{(s-1)\sum_{i=n+1}^{2n} T_i} > \frac{T_0}{T_n} \cdot \frac{1}{s-1} \quad \text{and}$$

$$\frac{\sum_{i=-n}^{1} T_i}{\sum_{i=-nr}^{-n-1} T_i} \geqslant \frac{\sum_{i=-n}^{1} T_i}{(r-1)\sum_{i=-2n}^{-n-1} T_i} > \frac{T_0}{T_{-n}} \cdot \frac{1}{r-1}. \tag{27}$$

He can show that for every natural number

$$n \geqslant \max\left\{n_1 = m_1 + \frac{m_1 s - s}{r+1},\ n_2 = m_2 + \frac{m_2 r - r}{s+1}\right\}, \tag{28}$$

where m_1 and m_2 are the smallest natural numbers satisfying respectively

$$m_1 \geqslant \frac{\log[c(s-1)]}{\log(r+1)-\log r} \quad \text{and} \quad m_2 \geqslant \frac{\log[c(r-1)]}{\log(s+1)-\log s}, \tag{29}$$

that

$$\frac{T_0}{T_n} \geqslant c(s-1) \quad \text{and} \quad \frac{T_0}{T_{-n}} \geqslant c(r-1). \tag{30}$$

With this last step Bernoulli had all the elements in order to show that for every n satisfying the above mentioned condition

$$\frac{\Pr\{|h_{nt}-p| \leqslant 1/t\}}{\Pr\{|h_{nt}-p| > 1/t\}} = \frac{\sum_{i=-n}^{n} T_i}{\left[\sum_{i=-nr}^{-n-1} T_i + \sum_{i=n+1}^{ns} T_i\right]} > c \tag{31}$$

holds, or in his words the following proposition, his 'theorema aureum', is valid (p. 236). He begins it with an explanation:

> Finally follows the proposition for which all this has been said. Its demonstration shall be established only by the application of the aforementioned lemmas to the present situation. In order to avoid long circumscriptions I shall call those cases, in which a certain event can occur, favourable or fertile and sterile those, in which the certain event cannot occur. In the same way [I call] those experiments favourable or fertile in which one of the fertile cases is seen to happen, and unfavourable or sterile those in which one of the sterile cases is observed to happen. Be therefore the number of fertile to the number of sterile cases in the proportion of r/s and so to the number of all in the proportion of $r/(r+s)$ or r/t, which proportion is contained in the limits $(r+1)/t$ and $(r-1)/t$. It has to be shown that one can make so many experiments that it is a given number of times, say c-times, more probable for the number of fertile observations to fall within these limits than outside, which means that the proportion of the number of fertile to the number of all observations shall not be greater than $(r+1)/t$ and not be smaller than $(r-1)/t$.

In an appendix Bernoulli treats the *jeu de paume* as a game of chance by taking the relative frequency of winning as a measure of the probability of winning a single game.

The title 'Ars conjectandi' was suggested by the *Ars cogitandi*, better known as the Logic of Port Royal, in the very last chapter of which the chances for future contingent events are equated with the ratios of the associated degrees of probability. One can see how Bernoulli, beginning from this notion, developed the classical concept of probability, and how he became the first to set down the prerequisites for consciously formulating a programme for the mathematization of all the fields of application subject to 'probabilis'. The fact that he had no time to illustrate his art of conjecturing by some examples from the social domain constituted a task that was taken up seriously only generations later.

7 THE IMPACT OF THE *ARS CONJECTANDI*

Jakob Hermann, a former student of Jakob Bernoulli who was trusted by the family, informed the authors of eloges which appeared in honour of Jakob Bernoulli about the content of the *Ars conjectandi* and especially the law of large numbers. It can be shown that already these rather short pieces of information influenced Montmort and de Moivre, who learned from them that the main concept around which the new doctrine of chances should be built is that of probability.

Nikolaus Bernoulli, Jakob's nephew, had tried to apply the findings of his uncle to a series of concrete problems in law in his dissertation *De usu artis conjectandi in jure* (1709). Montmort included part of his extensive correspondence with Nikolaus Bernoulli on the new field in the second edition of his *Essay d'analyse sur les jeux des hazard*, which appeared late in 1713. Both agreed in their judgement that the *Ars conjectandi* appeared too late to offer something new for specialists—understandably in the light of their own efforts to develop the subject. Their judgement has been interpreted as a proof of the lack of impact of Jakob's work, though both had been beneficiaries of Jakob's results contained in the *Ars conjectandi*. For Jakob's main theorem, the law of large numbers, Nikolaus Bernoulli offered a modified proof. Jakob's law of large numbers stimulated de Moivre, who late in his life found the first form of the central limit theorem that allows the approximation of the binomial by the normal distribution (§7.6).

Bernoulli's program to mathematize the realm of the probable, including what now is called the social domain, occupied mathematicians throughout the 18th and 19th century until the advent of mathematical statistics (§56). The peak of this development was reached in the late Enlightenment when the Marquis de Condorcet and P.S. Laplace held the view (§24) that the science of the probable is the only reliable guide to decide questions where we have no certain knowledge like in things concerning the outcome of votes or the reliability of witnesses.

BIBLIOGRAPHY

Edwards, A.W.F. 1987. *Pascal's arithmetical triangle*, London: Oxford University Press. [Repr. Baltimore: Johns Hopkins University Press, 2002.]

Hacking, I. 1975. *The emergence of probability*, Cambridge: Cambridge University Press.

Hald, A. 1990. *A history of probability and statistics and their applications before 1750*, New York: Wiley.

Schneider, I. 1981. 'Why do we find the origin of a calculus of probabilities in the seventeenth century?', in J. Hintikka et alii (eds.), *Probabilistic thinking, thermodynamics, and the interaction of the history and philosophy of science*, vol. 2, Dordrecht: Reidel, 3–24.

Schneider, I. 1984. 'The role of Leibniz and of Jakob Bernoulli for the development of probability theory', *Llull, 7*, 68–89.

Schneider, I. 1993. *Johannes Faulhaber (1580–1635)—Rechenmeister in einer Welt des Umbruchs*, Basel: Birkhäuser (*Vita mathematica*, vol. 7).

Schneider, I. 2001. 'Jakob Bernoulli', in Chris Heyde and Eugene Senata (eds.), *Statisticians of the centuries*, New York: Springer, 33–38.

Shafer, G. 1978. 'Non-additive probabilities in the work of Bernoulli and Lambert', *Archive for history of exact sciences, 19*, 309–370.

Stigler, S.M. 1986. *The history of statistics. The measurement of uncertainty before 1900*, Cambridge, MA: Harvard University Press.

Todhunter, I. 1865. *A history of the mathematical theory of probability from the time of Pascal to that of Laplace*, Cambridge: Macmillans. [Repr. New York: Chelsea, 1949.]

CHAPTER 7

ABRAHAM DE MOIVRE, *THE DOCTRINE OF CHANCES* (1718, 1738, 1756)

Ivo Schneider

The *Doctrine of chances* is the first textbook for the calculus of probabilities. It constitutes the results of the activities of its author as a private instructor of mathematics. It was based on the concept of probability and its classical measure; it contained in an introductory theoretical part the main rules, extended the mathematical methods for the solution of its problems by analytical tools, and offered from the second edition on an approximation of the binomial by the normal distribution.

First publication. The doctrine of chances: or, a method for calculating the probability of events in play, London: W. Pearson, 1718. xiv + 175 pages.

Second edition. London: Woodfall, 1738. xiv + 258 pages. [Photorepr. London: Cass, 1967.]

Third edition (posthumous). London: A. Millar, 1756. xii + 348 pages. [Photorepr. New York, Chelsea, 1967.]

Partial Italian translation. Abramo Moivre la dottrina degli azzardi aplicata ai problemi della probabilità della vita, delle pensioni vitalizie, reversioni, tontine, ec. transportata dall' idioma Inglese, arrichita di note ed aggiunte [. . .] (trans. P. Don Roberto Gaeta with P. Don Gregorio Fontana), Milan: Galeazzi, 1776. [Translates (De Moivre 1725) on annuities.]

Partial French translation by P. Crépel in Thierry Martin (ed.), *L'Arithmétique politique en France au XVIIIe siècle*, Paris: INED, 2004, to appear. [Of the preliminary discourse.]

Related articles: Jakob Bernoulli (§6), Bayes (§15), Laplace on probability (§24).

1 BACKGROUND AND STORY OF THE PUBLICATION

Abraham Moivre, born on 26 May 1667 as the son of a Protestant surgeon in Vitry-le-François in the Champagne region of France, spent the first 20 years of his life in France,

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

where he was educated in different Huguenot institutions. Presumably influenced by René Descartes, he had turned to mathematics. By the age of 16 years he had studied amongst other things the tract 'De ratiociniis in ludo aleae' of the Cartesian Christiaan Huygens. In Paris in the 1680s he was taught mathematics by the private teacher of mathematics Jacques Ozanam, who might have been seen by Moivre as a model for making his living when he had to support himself shortly afterwards.

After the revocation of the Edict of Nantes in 1685 hundreds of thousands of Huguenots left France. Amongst them was Moivre, who went to England where, together with his brother Daniel, he was granted denization in December 1687 and naturalisation in 1705. In England he began his occupation as a tutor in mathematics. Here he added a 'de' to his name. De Moivre mastered Isaac Newton's *Principia* (1687) (§5) very early and became a true and loyal Newtonian. Trying to make a reputation as a mathematician in the new field of the infinitesimal calculus, he had a very unpleasant fight with George Cheyne between 1703 and 1705; further, impressed by the superiority of Johann Bernoulli, with whom he had corresponded between 1704 and 1714, he decided to leave analysis. He then engaged in the calculus of games of chance and probability theory, which was of great interest for many of his students and where he hoped for fewer competitors. On de Moivre's life and work see [Schneider, 1968].

Next to clients like Francis Robartes it was P.R. de Montmort (1678–1719) who had raised de Moivre's interest in the theory of games of chance and probability, with the first edition of his *Essay d'analyse sur les jeux de hazard* (1708). De Moivre published an article 'De mensura sortis' on the subject in the *Philosophical Transactions* of the Royal Society [De Moivre, 1712], to which he had been elected in 1697. The paper was followed by his *Doctrine of chances (*hereafter, '*DoC')*, published in 1718. A second edition (1738) contained his normal approximation to the binomial distribution, which he had found in 1733. The third edition (1756) contained as a second part his *A treatise of annuities on lives*, that had been published as a monograph in 1725.

The *DoC* is in part the result of a competition between de Moivre on the one hand and Montmort together with Nikolaus Bernoulli on the other. De Moivre claimed that his representation of the solutions of the then current problems tended to be more general than those of Montmort, which Montmort resented very much. This situation led to some arguments between the two men, which finally were resolved by Montmort's premature death in 1719. His grievances are understandable in the light of what de Moivre had achieved with the 'De mensura sortis' and the first edition of *DoC*, even if the first edition already went beyond some of the results achieved by Montmort. However, the second and third editions of the *DoC* offered so many new results that Montmort's contribution justly fell into oblivion. De Moivre had developed algebraic and analytical tools for the theory of probability like a 'new algebra' for the solution of the problem of coincidences which somewhat foreshadowed Boolean algebra (compare §36), and also the method of generating functions or the theory of recurrent series for the solution of difference equations (compare Laplace in §24). Differently from Montmort, de Moivre offered in *DoC* an introduction that contains the main concepts like probability, conditional probability, expectation, dependent and independent events, the multiplication rule, and the binomial distribution.

De Moivre's greatest mathematical achievement is considered a form of the central limit theorem, which he found in 1733 at the age of 66. He understood his central limit theorem

as a generalization and a sharpening of Bernoulli's *Theorema aureum* (§6.6), which was later named 'the law of large numbers' by S.D. Poisson.

2 THE INTRODUCTION OF THE *DOCTRINE OF CHANCES*, AND THE MATHEMATICAL REQUIREMENTS FOR THE READER AS ANNOUNCED IN THE PREFACE

The contents of the first and third editions of the book are summarised in Tables 1 and 2; they give some idea of the scale of the changes. The survey below takes note of all three editions. For general accounts of the contents and context see [Todhunter, 1865, ch. 9; David, 1962, ch. 15; Hald, 1990, chs. 19–25].

As is underlined in the introduction, the book has the character of a textbook for autodidactic study as well as a companion for private instruction: with its help the reader should 'be able to solve a great variety of questions depending on chance'. That there would remain sufficiently many unanswered questions, especially for those only versed in 'common arithmetick', is indicated by the following sentence: 'I wish [...] that I could every where have been as plain as in the Introduction; but this was hardly practicable'.

De Moivre begins with the classical measure of probability, 'a fraction whereof the numerator be the number of chances whereby an event may happen, and the denominator the number of alls the chances whereby it may either happen or fail'. He gives the summation rule for probabilities of disjunct events explicitly only for the case of the happening and the not happening of an event. Expectation is still on the level of Huygens defined as the product of an expected sum of money and the probability of obtaining it, the expectation of

Table 1. Contents by Problems of de Moivre's book (first edition, 1718). The Roman numerals refer to Problems.

Items	**Page**	**Topics**
Dedication		Dedication to Isaac Newton.
Preface	i–xiv	Outline of the book together with the mathematical prerequisites for its different parts.
Introduction	1	Addition and multiplication theorem for dependent and independent events; expectation.
I–XIV	9	Different problems solvable with the rules contained in the introduction, including ones dealing with the games of Bassette (XII) and Pharaon (XIII).
XV–XXXII	47	Problems solvable by combinatorial methods, including some dealing with lotteries (XXI and XXII), and of Pharaon (XXIII).
XXXIII–XLVI	102	The problem of the duration of play, or the ruin problem.
XLVII–LIII	155	Problems solvable by combinatorial methods: includes Hazard (XLVII, LIII), Whisk (XLVIII), Raffling (XLIX) and Piquet (LI, LII). [End 175.]

Table 2. Contents by Problems or Parts of de Moivre's book (third edition, 1756). The Roman numerals refer to Problems.

Items	**Page**	**Topics**
Dedication		Dedication to Lord Carpenter.
Preface	i	Outline of the book together with the mathematical prerequisites for the different parts of the book.
	xi	Advertisement.
Introduction	1	Addition and multiplication theorem for dependent and independent events; binomial distribution.
I–XIV	35	Different problems solvable with the rules contained in the introduction including problems dealing with the games of Bassette (XIII) and Pharaon (XIV).
XV–LVII	82	Problems solvable by with combinatorial methods including ones dealing with lotteries (XXI–XXV), the games of Quadrille (XXVII–XXXII), Pharaon (XXXIII), Hazard (XLVI, XLVII), Raffling (XLVIII, IL), Whisk (L), and Piquet (LI to LV).
LVIII–LXXI	191	The problem of the duration of play or the ruin problem.
LXXII–LXXIII	239	The deviation from the expected value.
	242	The approximation of the binomial by the normal distribution.
LXXIV	254	The probability of a run of given length.
	261	'Treatise of annuities on lives'; Preface.
Part I	265	Rules and examples covering Problems I–XXXIII, including single lives, reversions (271), successive lives (278), survivorship (282), expectation of life (288) and tables of the values of an annuity under different conditions (298).
Part II	310	Demonstrations of the main rules in Part I.
Appendices I–VII	329	Dedication of the first edition (1718) to Isaac Newton; notes to problems VII (app. I), IX (II), and XLV (III); table of the sums of logarithms together with Stirling's formula for $n!$ (IV); reprint of [De Moivre, 1744] on annuities of lives (VI); tables for the 'probabilities' of human life from Halley, Kersseboom, Parcieux, Smart and Simpson (VII). [End 348.]

several sums is determined by the sum of the expectations of the singular sums. He defines independent and dependent events and gives the multiplication rule for both. But whereas today the criterion for independence of two events is the validity of the multiplication rule in the *DoC*, the multiplication rule follows from the independence of the events, which seems to be a self-evident concept for de Moivre. By a series of concrete cases he derives the probability that an event with probability p happens in the $(l+1)$th trial and $(l-1)$

times in the trials before, where $i = 0, 1, \ldots, n - l$, as

$$p^l \sum_{i=0}^{n-l} \binom{l-1+i}{i} (1-p)^i, \tag{1}$$

which is based on the distribution

$$\binom{l-1+i}{i} p^l (1-p)^i = \binom{l-1+i}{l-1} p^l (1-p)^i, \tag{2}$$

today identified with the negative binomial distribution. Later he applies this solution to the problem of points (on the division of the stake between the players when a game is interrupted before the end).

In a similar way de Moivre derives the probability that an event with probability p happens at least l times in n (independent) trials as

$$\sum_{i=0}^{n-l} \binom{n}{i} p^{n-i} (1-p)^i. \tag{3}$$

Accordingly he gives the binomial distribution as the probability that an event with probability p happens exactly l times in n (independent) trials:

$$\binom{n}{l} p^l (1-p)^{n-l}. \tag{4}$$

With these tools 'those who are acquainted with Arithmetical Operations' (as de Moivre remarked in the preface) could tackle many problems, in part already well known but which he gradually generalized. Because the majority of the solved problems depends on rules 'being entirely owing to Algebra' and to combinatorics, de Moivre tried to convince those readers who had not studied algebra yet to 'take the small Pains of being acquainted with the bare Notation of Algebra, which might be done in the hundredth part of the Time that is spent in learning to write Short-hand'. Remarks of this kind are typical of the private teacher of mathematics de Moivre, who was accustomed to ask his clients before he began with his instructions about their mathematical knowledge.

Some problems, as already stated by Jakob Bernoulli (1654–1705) in his *Ars conjectandi* (§6), can be solved more easily by the use of infinite series. As an illustration de Moivre offers the problem to determine the amounts each of two players A and B has to stake under the condition that the player who throws the first time an Ace with an ordinary die wins the stake and that A has the first throw. He considers it as reasonable that A should pay $\frac{1}{6}$ of the total stake in order to have the first throw, B should pay $\frac{1}{6}$ of the rest which is $\frac{1}{6} \cdot \frac{5}{6}$ for having the second throw, A should pay $\frac{1}{6}$ of the remainder for having the third throw, etc. The part that A has to stake altogether is the sum of a geometrical series with $\frac{1}{6}$ as the first term and the quotient $\frac{25}{36}$, which is $\frac{6}{11}$ of the total stake. Accordingly B's share is $\frac{5}{11}$ of the total stake. De Moivre claims that in most cases where the solution affords the application of infinite series the series are geometrical. The other kind of infinite series

which relate to the problem of the duration of play are recurrent series the terms of which can be connected with the terms of geometrical series, as will be explained later. Other problems depend on the summation of the terms of arithmetical series of higher orders and a 'new sort of algebra'.

It seems appropriate to present now the main results contained in the *DoC* following the methods applied for the solution of its problems.

3 GENERATING FUNCTIONS

The problem that induced de Moivre to introduce what was later called by P.S. Laplace (1749–1827) a 'generating function' (§24.4) is a lemma of problem III. It asks after the number of chances to throw a given number $p+1$ of points with n dice, each of them of the same number f of faces. Here the word 'dice' or 'die' is used in the more general sense of, for example, a roulette wheel with f sectors.

The exponents of the terms of the function

$$f(r) = 1 + r + rr + r^3 + \cdots + r^{f-1} = \frac{1 - r^f}{1 - r} \tag{5}$$

represent the f different 'faces' with the number of points diminished by 1 on each face. The coefficient of the term with the exponent $p+1-n$ in the development of the function

$$g(r) = \big(f(r)\big)^n = \big(1 - r^f\big)^n \cdot (1 - r)^{-n} = \left(\sum_{i=0}^{n} \binom{n}{i} (-1)^i r^{if}\right)\left(\sum_{j \geqslant 0} \binom{n+j-1}{j} r^j\right), \tag{6}$$

which is because

$$p + 1 - n = if + j, \quad i = 0, 1, \ldots, \frac{p+1-n}{f}, \tag{7}$$

and

$$\binom{n+j-1}{j} = \binom{p-if}{p-if-n+1} = \binom{p-if}{n-1} = \sum_{i=0}^{(p+1-n)/f} (-1)^i \binom{n}{i} \binom{p-if}{n-1}. \tag{8}$$

So the function $g(r)$ generates the solution or is the generating function of the number of chances sought after.

4 A 'NEW SORT OF ALGEBRA' IN THE *DOCTRINE OF CHANCES*

In the preface of the 1756 *DoC*, which is a shortened version of the preface from 1718, de Moivre remarks: 'In the 35th and 36th Problems, I explain a new sort of Algebra, whereby some Questions relating to Combinations are solved by so easy a Process that their solution is made in some measure an immediate consequence of the Method of Notation' (p. ix). He conceded later that one could in any case also reach the solution with comparatively

less simple and general methods. The use of plus and minus signs for combining probabilities, as an extension from their normal use for real magnitudes, was a decisive step in the evolution of what he declared with some pride to be his 'new sort of Algebra'.

In this respect de Moivre met with a much more receptive mathematical environment when one recalls the objections and resistance to the use of an algebraic calculus for Aristotelian logic that continued right up to the middle of the 19th century. On the other hand De Moivre did not create a new logical calculus nor, like George Boole in his *An investigation of the laws of thought* (1854), did he establish the requirement for probability theory to conform to set algebra (compare §36.5).

The idea behind this new sort of algebra is illustrated by de Moivre's solution of the following problem: find the probability that, of the n first letters of the alphabet in some order, m should be in their original position, l should not be in their original position and the remaining $n-(m+l)$ in arbitrary positions. He proceeded inductively beginning with the most simple cases by using the following notation (pp. 111 f.):

> [...] so that, for instance $a+b+c-d-e$, may denote the probability that a, b, and c shall be in their proper places, and that at the same time both d and e shall be excluded their proper places.
> It having been demonstrated in what we have said of Permutations and Combinations, that $a=\frac{1}{n}$; $a+b=\frac{1}{n\cdot(n-1)}$; $a+b+c=\frac{1}{n\cdot(n-1)\cdot(n-2)}$, &c. let $\frac{1}{n}$, $\frac{1}{n\cdot(n-1)}$, &c. be respectively called r, s, t, v, &c. this being supposed, we may come to the following Conclusions.
>
> $b=r$,
>
> $b+a=s$,
>
> then
>
> 1°. $\underline{b-a=r-s}$
>
> $c+b=s$ for the same reason that $a+b=s$.
>
> $\underline{c+b+a=t}$.
>
> 2°. $\underline{c+b-a=s-t}$
>
> $c-a=r-s$ by the first Conclusion.
>
> $\underline{c-a+b=+s-t}$ by the second.
>
> 3°. $\underline{c-a-b=r-2s+t}$
>
> $d+c+b=t$.
>
> $\underline{d+c+b+a=v}$.
>
> 4°. $\underline{d+c+b-a=t-v}$
>
> $d+c-a=s-t$ by the second Conclusion.
>
> $\underline{d+c-a+b}=t-v$ by the fourth.

5°. $\underline{d+c-a-b} = s - 2t + v$

$d - b - a = r - 2s + t$ by the third Conclusion.

$\underline{d-b-a+c} = s - 2t + v$ by the fifth.

6°. $\underline{d-b-a-c} = r - 3s + 3t - v$.

Here he is using the plus and minus signs on the left-hand side to stand for logical *and* and *and not*, while the same signs on the right-hand side stand for the usual addition and subtraction of real numbers. De Moivre's understanding of algebra as a collection of mathematical objects, associated with operations for combining them agrees, of course, with ordinary literal algebra.

5 THE DURATION OF PLAY

One of the problems with a long tradition is that of the duration of play. It resulted from a generalization of the last problem that Huygens had posed to his readers at the end of his treatise *De ratiociniis in ludo aleae* (1656). The first to deal with the problem in the new form seems to be Montmort, and after him Nikolaus Bernoulli. De Moivre concerned himself with it at about the same time. His formulation of the problem in the *DoC* of 1718 is nearly the same as he used in the third edition (p. 191):

> Two gamesters A and B whose proportion of skill is as a to b, each having a certain number of pieces, play together on condition that as often as A wins a game, B shall give him one piece; and that as often as B wins a game, A shall give him one piece; and that they cease not to play till such time as either one or the other has got all the pieces of his adversary: now let us suppose two spectators R and S concerning themselves about the ending of the play, the first of them laying that the play will be ended in a certain number of games which he assigns, the other laying to the contrary. To find the probability that S has of winning his wager.

De Moivre solves the problem step by step. He begins with the assumption that both A and B have the same number n of pieces and that $n + d$ games are played. Starting from $n = 2$ and $d = 0$ he derives an algorithm for the determination of the probability $p_{n+d}(S)$ that neither player is ruined after $n + d$ games. For the most simple case $n = 2$ and $d = 0$ one sees that from the events related to the terms of $(p+q)^2 = p^2 + 2pq + q^2$ only the term $2pq$ counts for S, or $p_{2+0}(S) = 2pq$. For $n = 2$ and $d = 1$ he gets $p_{2+1}(S) = 2pq$, that is to say the same value for the probability that neither player is ruined. In the general case his solution presupposes d even.

It is plausible from the development of

$$(p+q)^{n+d} = (p+q)^n (p+q)^d = (p+q)^n \left[(p+q)^2\right]^{d/2} \tag{9}$$

that $p_{n+d}(R)$, which is the sum of the terms representing the events that either A or B wins n times in the first n games, that either A or B wins $n + 1$ times in the first $n + 2$ games, ..., and that either A or B wins $(n + d/2)$ times in $n + d$ games, is of the form

$$p_{n+d}(R) = \left(p^n + q^n\right)\left[\sum_{\nu=0}^{[d/2]} \beta_{n,\nu}(pq)^\nu\right], \tag{10}$$

where the $\beta_{n,\nu}$ are the terms of what de Moivre called a 'recurrent series'. He had defined recurrent series in the following way (p. 220 ff.):

> I call that a recurring series which is so constituted, that having taken at pleasure any number of its terms, each following term shall be related to the same number of preceding terms, according to a constant law of relation, such as the following series
>
> $$\begin{array}{cccccccc} A & B & C & D & E & F & & \\ 1+ & 2x+ & 3xx+ & 10x^3+ & 34x^4+ & 97x^5, & \text{\&c.} & \end{array}$$
>
> in which the terms being respectively represented by the capitals A, B, C, D, &c. we shall have
>
> $$\begin{aligned} D &= 3Cx - 2Bxx + 5Ax^3, \\ E &= 3Dx - 2Cxx + 5Bx^3, \\ F &= 3Ex - 2Dxx + 5Cx^3, \\ &\text{\&c.} \end{aligned}$$
>
> Now the quantities $3x - 2xx + 5x^3$, taken together and connected with their proper signs, is what I call the index, or the scale of relation; and sometimes the bare coefficients $3 - 2 + 5$ are called the scale of relation.

For the $\beta_{n,\nu}$ de Moivre could give not only the scale of relation but also the values of the $\beta_{n,\nu}$ themselves, which he seems to have found earlier by induction.

De Moivre had competed with Montmort and Nikolaus Bernoulli over the determination of the general term of a recurring sequence and the sum of a recurrent series in the 1720s; but he had perfected his theory of recurrent series in his *Miscellanea analytica* of 1730. He repeated the main results in the second and third edition of the *DoC,* albeit without demonstrations.

The main idea for finding the general term a_n of a recurrent sequence $((a_\nu))$ with known scale of relation of r terms is to resolve the $((a_\nu))$ in a sum of r geometrical sequences $((b_{\rho,\nu}))$, $\rho = 1, 2, \ldots, r$, so that $a_n = \sum_{\rho=1}^r b_{\rho,n} = \sum_{\rho=1}^r b_{\rho,0} \cdot q_\rho^n$, where the $b_{\rho,0}$ are the first terms of the $((b_{\rho,\nu}))$ and the q_ρ the respective quotients. The realization of this idea amounts in modern terms to the solution of a homogeneous linear difference equation with constant coefficients by introducing the characteristic equation of the difference equation. The r roots of the characteristic equation are the sought after quotients q_ρ, $\rho = 1, 2, \ldots, r$, under the condition that these roots are real and different from one another, which is not specified by de Moivre. The first terms of the geometrical sequences are the solutions of

the inhomogeneous system of r linear equations

$$a_\nu = \sum_{\rho=1}^{r} b_{\rho,0} \cdot q_\rho^\nu, \quad \nu = 0, \ldots, r-1. \tag{11}$$

Already in the first edition of *DoC* de Moivre had given a solution for

$$p_{n+d}(S) = 1 - p_{n+d}(R) = \sum_{i=1}^{n-1} \alpha_{n,d/2,i} \cdot p^{n+d/2-i} q^{d/2+i}, \tag{12}$$

where the $\alpha_{n,d/2,i}$, $i = 1, \ldots, n-1$, are again a recurrent sequence, in the form of a trigonometric formula. His competitors were not able to explain it since he had never revealed his derivation. Obviously he had seen that the coefficients of the respective characteristic equation for the special case $p = q = 1/2$ and n even are the same as in the equation

$$2\cos n\alpha = \sum_{i=0}^{n/2} a_{2i} (2\cos\alpha)^{n-2i}, \tag{13}$$

with which he was familiar long before he published it in the *Philosophical transactions* for 1738. He was so proud of this trigonometric representation for the solution of the problem of the duration of play that he indicated it on a vignette showing the Goddess Fortuna with a geometrical representation of it. It appears on the title page of all editions of the *DoC*; Figure 1 shows that for the first edition.

Already in 'De mensura sortis' de Moivre had generalized the problem of the duration of play to the case where the number of pieces that the two players A and B hold at the beginning are different. However, from a methodological point of view the solution of this generalized problem offers nothing new. In the paper he had also shown that the coefficients α_i in the equation

$$p_{n+d}(S) = \sum_{i=1}^{n-1} \alpha_i \cdot p^{n+d/2-i} q^{d/2+i} = c \quad \text{satisfy } \alpha_i = \alpha_{n-i}, \tag{14}$$

and that such an equation of degree $n + d$ can be reduced by an appropriate substitution to an equation of degree $(n + d)/2$. He used this property in the *DoC* in order to solve problems of this kind: given the number of pieces each player has in the beginning, 'what must be their proportion of Chances for winning any one Game assigned, to make it as probable that the Play will be ended in a certain number of games as not?'.

6 THE APPROXIMATION OF THE BINOMIAL BY THE NORMAL DISTRIBUTION

In modern terms de Moivre was interested, in a way different from Jakob Bernoulli, in the determination of ε as a function of n and d in the equation:

$$P\big(|h_n - p| \leqslant \varepsilon\big) = d, \tag{15}$$

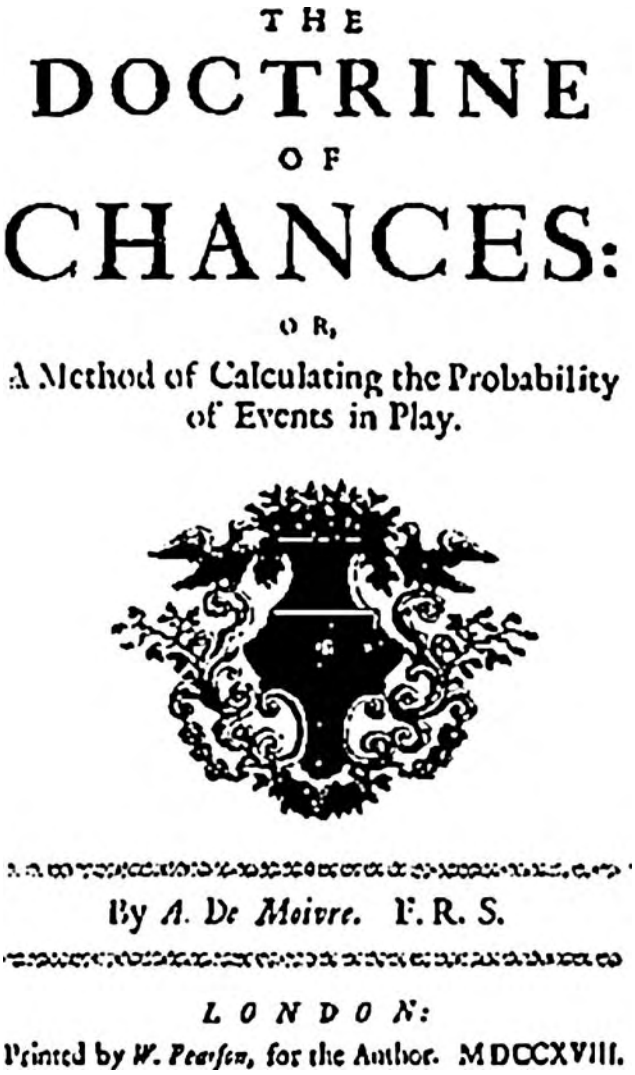

THE
DOCTRINE
OF
CHANCES:
OR,
A Method of Calculating the Probability
of Events in Play.
By A. De Moivre. F. R. S.
LONDON:
Printed by W. Pearson, for the Author. MDCCXVIII.

Figure 1. Title page of De Moivre's first edition.

where P signifies the probability that the relative frequency h_n of the appearances of an event in n independent trials does not deviate from its expected value p by more than ε. He started with the simplest case the symmetric binomial ($p = 1/2$) and $\varepsilon = l/n$. First he estimated the maximum term $b(m)$ in the sum

$$\sum_{k=m-l}^{m+l} \binom{n}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{n-k} = 2^{-n} \sum_{k=m-l}^{m+l} \binom{n}{k} =: \sum_{i=-l}^{l} b(m+i) \tag{16}$$

for large n $(= 2m)$ as

$$b(m) \cong \frac{2}{\sqrt{2\pi n}}. \tag{17}$$

In a second step he considered the ratio of a term $b(m-l)$ with a distance l, $O(\sqrt{n})$, from the maximum to the maximum, for which he found for large n:

$$\ln \frac{b(m \pm l)}{b(m)} \cong -\frac{2l^2}{n}. \tag{18}$$

De Moivre had worked on these estimations between 1721 and 1733; important for the final form of these estimations were asymptotic series for $n!$ with n large, to which both he and James Stirling had contributed and published in 1730. The divergence of these series was explicitly denied by de Moivre; their asymptotic behaviour was recognized only much later.

With these estimations de Moivre could give the following approximation for large n:

$$P\left(\left|h_n - \frac{1}{2}\right| \leqslant \frac{l}{n}\right) = 2^{-n} \sum_{k=m-l}^{m+l} \binom{n}{k} \cong \frac{4}{\sqrt{2\pi n}} \sum_{k=0}^{l} e^{-2k^2/n} \cong \frac{4}{\sqrt{2\pi n}} \int_{x=0}^{l} e^{-2x^2/n}\, dx. \tag{19}$$

He did not use this form of representation; especially the last integral was represented, typical for the Newtonian school, in form of the series

$$\frac{4}{\sqrt{2\pi n}} \sum_{i=0}^{\infty} \frac{(-1)^i 2^i l^{2i+1}}{i!(2i+1)n^i}. \tag{20}$$

He saw that this series converges numerically very fast for $l = \sqrt{n}/2$ and calculated its values for $l = s\sqrt{n}/2$ with s 1, 2 and 3. In addition, he gave the corresponding estimations for the general binomial (for any p). Schneider [1996] had shown that these estimations lead to the series

$$\frac{2}{\sqrt{2\pi pqn}} \sum_{i=0}^{\infty} \frac{(-1)^i l^{2i+1}}{i!(2i+1)(2pqn)^i} \quad \text{for } P\left(|h_n - p| \leqslant \frac{l}{n}\right), \tag{21}$$

which for $l = s\sqrt{npq}$ is the same as the one found by de Moivre for the case of $p = 1/2$. All this demonstrates that he understood intuitively the importance of what was later called the standard deviation.

In this way de Moivre could show with his approximation of the binomial through the normal distribution, which he used in order to avoid the tedious calculations of the coefficients of the binomial distribution, that for large n and an $\varepsilon = s\frac{\sqrt{npq}}{n}$ the probability $P(|h_n - p| \leqslant \varepsilon)$ is approximately $0.684\ldots$ for $s = 1$, $0.954\ldots$ for $s = 2$, and $0.998\ldots$ for $s = 3$. The approximation of the binomial through the normal distribution with its consequences was the culmination of the *DoC* from the second edition onwards. He used this purely mathematical result as a means in order to combine the theory of probability with natural religion. Most important for such an attempt was his interpretation of the terms

probability and chance. Chance had for him two connotations. In the first remark to his central limit theorem he used the term 'chance' as an antithesis to law, here statistical law. The existence of laws of this kind is due to a 'design' which according to contemporary convictions relates immediately to divine providence. Chance in contrast appears as something which obscures this design by 'irregularity'. This had to do with the unpredictability of the outcome in single trials.

In a second remark on the central limit theorem, which appeared only in the third edition of the *DoC*, de Moivre illustrated the relationship between chance and law, or design, by the example of the stability of the sex ratio of newborn in London for a period of more than 80 years. In an article in the *Philosophical transactions* for 1710 John Arbuthnot had chosen design as the determinant cause for the observed sex ratio effect; but in published letters to Montmort from 1712 and 1713 Nikolaus Bernoulli had voted for chance. Since the sex ratio for the whole period could be approximated very well by the ratio of $18 : 17$ in favour of boys, Bernoulli proposed to throw a 'die' of $18 + 17$ faces a number of times equal to the number of births. Armed with his calculations he could claim a very high probability for the fact that the ratio produced by throwing this die, that is, by chance, differs very little from the observed sex ratio.

De Moivre tried to combine the two positions by arguing that Bernoulli's 'die' had to be made by an artist who followed in the execution of it a plan, a design and thus not chance in Bernoulli's understanding. According to de Moivre, chance 'supposes the *Existence* of things, and their general known *Properties* that a number of Dice, for instance, being thrown, each of them shall settle upon one or other of its Bases'. Accordingly 'the *Probability* of an assigned Chance, that is of some particular disposition of the Dice, becomes as proper a subject of Investigation as any other quantity or Ratio can be'. Obviously this connotation of chance refers to the possible outcome of an event. Chance understood as irregularity and unpredictability in a small number of trials but not in the long run, is consistent with an all-wise and all-powerful creator who had not abandoned his creation after its perfection but rules it permanently in order to guarantee its existence. For de Moivre probabilities, interpreted as laws which express God's design, are objective properties of creation and chance, as an existing property of the material world with its irregular and unpredictable aspects, is a manifestation of God's constant involvement in the course of his creation and so is objective in the sense that it is independent of the human subject and its level of information.

7 ANNUITIES ON LIVES

The most elaborate version of de Moivre's *Annuities on lives* appeared in the third edition of the *DoC*. This shows already that he followed Jakob Bernoulli in treating questions concerning life insurance, especially annuities on lives as a legitimate part of probability theory. Nevertheless, he had originally published his work as an independent monograph [De Moivre, 1725]. His preoccupation with questions concerning interest, loan, mortgage, pensions, reversions or annuities goes back at least to the 1690s, from which time a piece of paper, kept in the *Autographensammlung* in the *Preussische Staatsbibliothek* in Berlin, contains de Moivre's answers to pertinent questions of a client.

At this time Halley had reconstructed from the lists of births and deaths in Breslau for each of the years 1687–1691 the demographic structure of the population of Breslau which he assumed as stationary in form of a life table. Halley's life table was published in the *Philosophical transactions* for 1693, together with applications to annuities on lives. Besides the formulas for the values of an annuity for a single life and for several lives he had calculated a table for the values of annuities of a single life for every fifth year of age at an interest rate of 6%. The immense amount of calculation work hindered him from doing the same for two and more lives. De Moivre solved this problem by a simplification. He replaced Halley's life table by a (piecewise) linear function. Based on such a hypothetical law of mortality and fixed rates of interest, he could derive formulas for annuities of single lives and approximations for annuities of joint lives as a function of the corresponding annuities on single lives. These results were published in his book *Annuities upon lives* (1725), together with the solution of problems of reversionary annuities, annuities on successive lives, tontines, and other contracts that depend on interest and the 'probability of the duration of life'.

In the second edition of the *DoC* part of the material contained in the *Annuities* together with new material was incorporated. After three more improved editions of the *Annuities* in 1743, 1750, and 1752 the last version of it was published in the third edition of the *DoC*. In it de Moivre begins with the supposition that the 'probabilities of life' 'decrease in arithmetical progression' which hypothesis compared with Halley's life table 'will be found to be exceedingly approaching'. The following 33 problems of the first part contain the solutions or approximate solutions of many contracts covering the whole realm where the probability of life is involved. Problem I asks for the value of an annuity A for a single life of given age i compared with an annuity P certain for $86 - i = n$ years at a given interest $r - 1$. $86 - i$ years are what de Moivre calls 'the complement of life', because according to the life tables available at his time life, notwithstanding that some people get older, ends in the average with 86. He gives in part I the solution $A = (1 - \frac{r}{n}P)/(r - 1)$ without any further explanation or demonstration. All he does is to repeat the solution in verbal form and to add a numerical example. In the same way he treats the other problems, some of which, such as those on survivorship or on the 'expectation of life', are not directly concerned with annuities. After the treatment of these 33 problems de Moivre includes tables for 'the present value of an annuity of one pound' certain for i years, $i = 1, \ldots, 100$, and tables for the value of an annuity for a single life of given age i according to the solution of problem I for interests at 3%, 3.5%, 4%, 5%, and 6%. The first part ends with four additional rules how to use the five tables for annuities certain for other purposes like finding the 'amount of the sum S in 7 years at $3\frac{1}{2}$ per Cent'.

The second part of the *Annuities on lives* in the *DoC* contained the 'demonstrations of some of the principal propositions in the foregoing treatise'. Here de Moivre distinguished 'real' lives corresponding to a piecewise linear approximation of a life table from fictitious lives where the probabilities to survive the next year or to die within the next year are constant independent of age. With the claim that 'the combination of two or more real lives will be very near the same as the combination of so many corresponding fictitious lives', and that accordingly the value of annuities on joint lives can be based on fictitious lives (p. 313), he can easily determine the value of annuities of joint lives as the sums of geometrical series. However, he concedes later that the transition from real to fictitious

lives, which he made for the sake of elegance and in order to avoid tedious calculations, 'creates an error too considerable to be neglected' (p. 327). So he ends the second part and with it the *DoC* with a 'General Rule for the Valuation of joint Lives' for the application of which he gives three numerical examples but no demonstration.

8 IMPACT OF THE *DoC*

An early reaction to the book which surely counts for the high estimation it was held at least in England is its exploitation by the Englishman Thomas Simpson, who in his *Treatise on the nature and laws of chance* (1740) just repeated the results achieved in the *DoC*. The fact that de Moivre had specialized in the theory of probability, for which he had prepared appropriate tools and to which he had contributed the solutions of the most interesting problems posed to him by his competitors and by his clients for some decades, made *DoC*, especially the last edition of 1756, the most complete representation of the new field in the second half of the 18th century.

This was felt by the leading mathematicians of the next generation. In particular, J.L. Lagrange and Laplace had planned a French translation of the book which however was never realized. Their interest goes back to de Moivre's solution of the problem of the duration of play by means of what he called 'recurrent series' and what amounts to the solution of a homogeneous linear difference equation with constant coefficients. In fact, the most effective analytical tool developed by Laplace for the calculus of probabilities, the theory of generating functions, is a consequence of his concern with recurrent series. Indeed, the most important results of the book reappear in Laplace's probability theory in a new mathematical form and in a new philosophical context (§24). This, more than anything else, confirms de Moivre's status as a pioneer in the field and as a predecessor of Laplace.

Another kind of impact was a side-effect of de Moivre's occupation with the ruin problem. Here he had found that the determination of the probability that the total game does not end after n single games leads to equations of the form

$$\sum_{\nu=0}^{n} a_\nu x^\nu = 0 \quad \text{with } a_\nu = a_{n-\nu}; \tag{22}$$

Leonhard Euler was to call these equations 'reciprocal'. De Moivre could show that for n even the equation could be reduced by a substitution equivalent to $x^2 + xy + 1 = 0$ to an equation of degree $n/2$, and that for odd n the equation has a root -1. It then follows that

$$\sum_{\nu=0}^{n} a_\nu x^\nu = (1+x)\sum_{\nu=0}^{n-1} b_\nu x^\nu = 0, \quad \text{where } \sum_{\nu=0}^{n-1} b_\nu x^\nu = 0, \tag{23}$$

is also a reciprocal equation. In this way, all reciprocal equations up to degree 9 can be solved, with solutions expressed in surds. These results concerning reciprocal equations were used by Euler to tackle a 'conjecture about the form of the roots of equations of any degree' of 1733.

The actuarial part of the *DoC*, the *Annuities on lives*, had perhaps an even longer impact. Whereas all his contemporaries depended for the determination of life expectancies on

mortality tables based on ever new observational material, de Moivre had propagated a law of mortality which can be described by a closed function, in his case a linear one. In the 18th century the first life insurance companies did not base their rates on his linear approach to mortality; however, in the 19th century Benjamin Gompertz found a nonlinear function, later improved by William Maitland Makeham, which resembled the data of the best mortality tables with sufficient accuracy. So de Moivre's idea of a mortality law found wide acceptance in life insurance mathematics.

BIBLIOGRAPHY

Anonymous. 1793. *Faro and Rouge et Noir: the mode of playing and explanation of the terms used at both games; with a table of the chances against the Punters extracted from De Moivre. To which is prefixed a history of cards*, London.

David, Florence Nightingale. 1962. *Games, gods and gambling*, London: Griffin.

De Moivre, A. 1712. 'De mensura sortis, seu, de probabilitate eventuum in ludis a casu fortuito pendentibus', *Philosophical transactions of the Royal Society of London*, (1711), 213–264.

De Moivre, A. 1725. *Annuities upon lives*, London. [Repr. in *DoC*, 3rd ed. (1756), 261–328.]

De Moivre, A. 1744. 'A letter from Mr. Abraham De Moivre, F.R.S. to William Jones, Esquire, F.R.S. concerning the easiest method for calculating the value of annuities upon lives, from tables of observations', *Philosophical transactions of the Royal Society of London*, 65–78.

Hald, A. 1990. *A history of probability and statistics and their applications before 1750*, New York: Wiley.

Schneider, I. 1968. 'Der Mathematiker Abraham de Moivre (1667–1754)', *Archive for history of exact sciences, 5*, 177–317.

Schneider, I. 1994. 'Abraham de Moivre: pionero de la teoría de probabilidades entre Jakob Bernoulli y Laplace', in E. de Bustos et alii (eds.), *Perspectivas actuales de lógica y filosofía de la ciencia*, Madrid: Siglo Veintiuno de España, 373–384.

Schneider, I. 1996. 'Die Rückführung des Allgemeinen auf den Sonderfall—eine Neubetrachtung des Grenzwertsatzes für binomiale Verteilungen von Abraham de Moivre', in J.W. Dauben et alii (eds.), *History of mathematics: states of the art*, San Diego: Academic Press, 263–275.

Stigler, S.M. 1986. *The history of statistics. The measurement of uncertainty before 1900*, Cambridge, MA: Harvard University Press.

Todhunter, I. 1865. *A history of the mathematical theory of probability from the time of Pascal to that of Laplace*, Cambridge: Macmillan. [Repr. New York: Chelsea, 1949.]

CHAPTER 8

GEORGE BERKELEY, *THE ANALYST* (1734)

D.M. Jesseph

The analyst is a criticism of the calculus, in both its Newtonian and Leibnizian formulations, arguing that the foundations of the calculus are incoherent and the reasoning employed in it is inconsistent. Berkeley's powerful objections provoked numerous responses, and the task of replying to them set the agenda for much of British mathematics in the 1730s and 1740s.

First publication. London: J. Tonson, 1734. viii + 94 pages.

Reprint. Dublin: S. Fuller and J. Leathley, 1734.

Reprints in editions of Berkeley's Works. 1) (ed. Joseph Stock), vol. 2, London: G. Robinson, 1784. 2) (ed. G.N. Wright), vol. 2, London: Tegg, 1843. 3) (ed. A.C. Fraser), vol. 3, Oxford: Clarendon Press, 1871 (repr. 1901). 4) (ed. George Sampson), vol. 3, London: G. Bell, 1898. 5) (ed. A.A. Luce and T.E. Jessop), vol. 4, London and New York: Nelson, 1951.

Other editions. 1) Berkeley, *De motu and the analyst* (ed. D.M. Jesseph), Dordrecht and Boston: Kluwer, 1992. 2) In P. Ewald (ed.), *From Kant to Hilbert*, vol. 1, Oxford: Clarendon Press, 1996, 60–92.

German translation. In Berkeley, *Schriften über die Grundlagen der Mathematik und Physik* (ed. and trans. W. Breidert), Frankfurt: Suhrkamp, 1969, 81–141.

French translations. 1) *L'analyste* (trans. A Leroy), Paris: Payot, 1936. 2) *L'analyste* (trans. J. Pignet), Paris: Centre de Documentation Universitaire, 1936. 3) *L'analyste* (ed. and trans. Michel Blay), in Berkeley, *Oeuvres* (ed. Geneviève Brykman), vol. 2, Paris: P.U.F., 1987, 257–332.

Related articles: Newton (§5), Leibniz (§4), MacLaurin (§10), Euler on the calculus (§14).

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

1 BERKELEY'S LIFE AND WORKS

George Berkeley (1685–1753) is best known for his contributions in philosophy, and more specifically for denying the existence of matter and propounding the idealistic thesis that only minds and ideas exist, so that in the case of 'sensible objects' *esse est percipi*, or to be is to be perceived. With the publication of his far-reaching critique of the foundations of the calculus in *The analyst*, he also established a name for himself in mathematics. In fact, the historian of mathematics Florian Cajori called the publication of *The analyst* 'the most spectacular event of the century in British mathematics' because of the impact of Berkeley's criticisms [Cajori, 1919, 57].

Berkeley was born in March of 1685 in the village of Thomastown in county Kilkenny, Ireland. He was educated at Trinity College, Dublin, where he took the degree B.A. in 1704 and was awarded a fellowship at Trinity upon taking the degree M.A. in 1707. His interest in mathematics was evident from early on, leading to his first publication, a small collection of mathematical works entitled *Arithmetica et miscellanea mathematica*.

Mathematics was not, however, the principal focus of Berkeley's intellectual efforts over the next few years. Between 1709 to 1713 he published three famous works, on which his philosophical reputation rests. His 1709 *Essay toward a new theory of vision* offered an account of perceived distance and magnitude in which the mind infers the properties of tangible objects on the basis of visual cues, but without perceiving properties like distance immediately. His *Treatise concerning the principles of human knowledge* of 1710 set out his idealistic philosophy in detail, arguing that the concept of 'material substance' is at once absurd and explanatorily useless. He pointed out that even philosophers who posit the existence of material bodies cannot explain how matter can produce ideas in the mind, or how purely mental phenomena like ideas could resemble or correspond to non-mental, material substances. Perhaps his most shocking claim in favor of his metaphysics was his oft-repeated contention that his principles were in strict accord with common sense and inimical to skepticism.

The *Principles* failed to be the success for which Berkeley had hoped. Among those who read it the consensus opinion was that his conclusions could not be seriously maintained and that his ultimate goal was either to promote skepticism or show his wit by propounding paradoxes. Indeed, a prominent London physician, upon reading the *Principles* concluded that Berkeley was mad and prescribed remedies for his affliction. Berkeley re-cast the fundamental tenets of the *Principles* as *Three dialogues between Hylas and Philonous* (1713), with the hope of presenting his system in a style more accessible to the general public.

This second formulation of Berkeley's philosophy was better received, but he still failed to generate any serious support for his system. He relocated from Dublin to London in 1713, remaining there until 1721 with interruptions for two tours of the Continent. In 1720, toward the end of his second Continental tour, he composed the small treatise *De motu* which he submitted to the Paris *Académie des Sciences* in competition for a prize on the nature and communication of motion. The essay applied Berkeley's strict empiricist epistemology and idealist metaphysics to the consideration of motion, concluding that terms like *force* lack empirical content except to the extent that they can be translated into talk about actually or possibly observed motions.

Berkeley established a literary reputation during his London years, associating with such figures as Jonathan Swift, Alexander Pope, Joseph Addison, and Sir Richard Steele, and contributing essays to Steele's periodical *The Guardian*. When he returned to Ireland in 1721 Berkeley had hopes of obtaining a preferment in the Church of Ireland, although complex political and theological factors kept him from realizing this ambition until 1724, when he was appointed Dean of Derry. This sinecure offered Berkeley the income and influence to campaign to found a college in America, a project in which he had been interested for some time. He published *A Proposal for the better supplying of churches in our foreign plantations, and for converting the savage Americans to Christianity* in 1724, in which he laid out the vision of a college in Bermuda he hoped to fund with private subscriptions as well as support from Parliament. In 1726 Parliament approved a £20,000 grant for the scheme, but delayed paying out the funds in the face of significant opposition. Berkeley set sail for Newport, Rhode Island in 1728 in the hope of encouraging Parliament to act, and he remained in Newport for three years until it became clear that the dream of founding a college in America had failed. While in Newport he wrote a collection of dialogues entitled *Alciphron, or the minute philosopher* in which he defended Christian doctrine against the claims of freethinkers, or, as he styled them, 'minute philosophers'.

After his return to England in 1731 Berkeley busied himself with the publication of *Alciphron* and other efforts in defense of Christianity, of which *The analyst* was a significant part. He also sought advancement in the Church, and in 1734 was appointed Bishop of Cloyne, near Cork. After assuming the bishopric in Cloyne, he made few notable philosophical, mathematical, or literary contributions.

2 THE PURPOSE OF *THE ANALYST*

Berkeley's publication of *The analyst* took up two themes that had long been of concern to him, one mathematical and the other theological. Mathematically, it continued the reservations about the foundations of the calculus that Berkeley had voiced in an early essay 'Of infinites' that he presented to the Dublin Philosophical Society in 1709 and reprised in arts. 130–132 of the *Principles*. Theologically, *The analyst* was part of Berkeley's battle against freethinking, and his principal argument intends to show that freethinkers who deride revealed religion for its mysteries cannot consistently accept the calculus, since it contains suppositions at least as extravagant and incomprehensible as anything in revealed religion. This aspect of his criticism is indicated in the full title of *The analyst*, which characterizes the work as *A Discourse addressed to an infidel mathematician; wherein it is examined whether the object, principles, and inferences of the modern analysis are more distinctly conceived or more evidently deduced, than religious mysteries and points of faith* and attributes it to 'The Author of *The minute philosopher*'.

Whether the work was directed at a specific 'infidel mathematician' is somewhat uncertain, although there is evidence that Berkeley intended it for Edmond Halley. According to Berkeley's 18th-century biographer Joseph Stock, the London physician Samuel Garth had declined the last rites in his final illness, on the grounds that 'my friend Dr. Halley who has dealt so much in demonstration has assured me that the doctrines of Christianity are incomprehensible and the religion itself an imposture'. According to Stock, Berkeley

'therefore took arms against this redoubtable dealer in demonstration, and addressed the *Analyst* to him, with a view of shewing, that Mysteries in Faith were unjustly objected to by mathematicians, who admitted much greater Mysteries, and even falsehoods in Science, of which he endeavoured to prove that the doctrine of fluxions furnished an eminent example' [Stock, 1776, 29–30]. Whomever *The analyst* was intended to address immediately, its broader audience was unmistakably those mathematicians who regarded the calculus as a rigorous and properly founded method that compared favorably with the mysterious tenets of revealed religion.

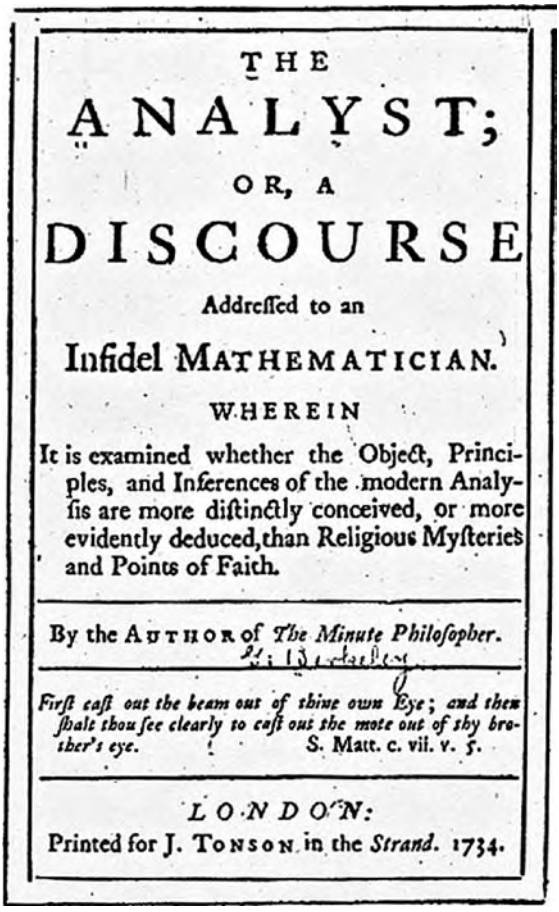
3 THE PRINCIPAL ARGUMENTS

The contents of Berkeley's book are summarised in Table 1, and its title page is Figure 1. His critique of the calculus contains two quite different sorts of criticisms. On the one hand, he argues that it violates canons of intelligibility by postulating incomprehensible entities such as infinitesimal magnitudes or ratios of evanescent quantities. On the other hand, he claims that the proofs of even the most elementary results in the calculus commit logical errors by employing inconsistent assumptions. We can distinguish these two sorts of arguments by classifying the former as metaphysical objections and the latter as logical objections.

The metaphysical criticism of the calculus proceeds by considering the definitions of its fundamental objects and asking whether these are, indeed, clear and comprehensible. It was a commonplace among philosophically minded mathematicians of the 18th century that their science dealt only with clearly grasped objects and sharply defined concepts

Table 1. Contents by Sections of Berkeley's book. viii + 94 pages.

Sections	**Description**
1–2	Introduction, contrasting religious mysteries with the principles of 'the modern analysis'.
3–8	The 'object of the calculus' considered and its foundational concepts dismissed as incomprehensible.
9–20	The 'principles and demonstrations' of the calculus critiqued, showing that both Newtonian and Leibnizian procedures seem to employ inconsistent assumptions.
21–29	Attempt to explain the reliability of the calculus as the result of its employing 'compensating errors' where a finite quantity is simultaneously overestimated and underestimated, while the errors balance precisely.
30–47	Various alternative formulations of the calculus considered and rejected, since all fail to overcome the fundamental objections.
48–50 and Queries	Conclusion and queries: a blanket indictment of the incomprehensible metaphysics and inconsistent reasoning of the 'modern analysts' with 67 'queries' ranging over various topics in mathematics and methodology.



THE

ANALYST;

OR, A

DISCOURSE

Addreſſed to an

Infidel MATHEMATICIAN.

WHEREIN

It is examined whether the Object, Principles, and Inferences of the modern Analyſis are more diſtinctly conceived, or more evidently deduced, than Religious Myſteries and Points of Faith.

By the AUTHOR of *The Minute Philoſopher*.

Firſt caſt out the beam out of thine own Eye; and then ſhalt thou ſee clearly to caſt out the mote out of thy brother's eye. S. Matt. c. vii. v. 5.

LONDON:

Printed for J. TONSON in the *Strand*. 1734.

Figure 1. Title page of Berkeley's book.

that exclude any obscurity or equivocation. To challenge this presumption, Berkeley begins with an inventory of the fundamental concepts of the Newtonian calculus of fluxions, specifically Newton's doctrine of moments and his definition of a fluxion as the velocity with which a geometric magnitude is produced (moving points producing lines, moving lines producing surfaces, etc.). He notes that moments are not finite particles, but rather 'the nascent Principles of finite Quantities', which are considered not to have any positive magnitude in themselves yet are capable of forming ratios to one another. Fluxions are not average velocities taken over a given time, but *instantaneous* velocities, defined as the ultimate ratios of evanascent increments of times and distances: 'These Fluxions are said to be nearly as the Increments of the flowing Quantities, generated in the least equal Particles of time; and to be accurately in the first Proportion of the nascent, or in the last of the evanescent, Increments' (Section 3). Moreover, higher-order fluxions are introduced by taking the fluxion itself as a variable quantity and considering the velocity with which it may change. Berkeley finds this whole apparatus inconceivable and ineradicably mysterious (Section 4):

> Now as our Sense is strained and puzzled with the perception of Objects extremely minute, even so the Imagination, which Faculty derives from Sense, is very much strained and puzzled to frame clear Ideas of the least Particles of time, or the least Increments generated therein: and much more so to comprehend the Moments, or those Increments of the flowing Quantites *in statu nascenti*, in their very first origin or beginning to exist, before they become finite particles. And it seems still more difficult, to conceive the abstacted Velocities of such nascent imperfect Entities. But the velocities of the Velocities, the second, third, fourth and fifth Velocities, &c. exceed, if I mistake not, all Humane Understanding. The further the mind analyseth and pursueth these fugitive Ideas, the more it is lost and bewildered; the Objects, at first fleeting and minute, soon vanishing out of sight.

The foundations of the Leibnizian *calculus differentialis* (§4) face a similar objection. Relying on the formulation of the Leibnizian differential calculus in L'Hôpital's *Analyse des infiniment petits* (1696), Berkeley claims that its postulation of infinitesimal magnitudes is incomprehensible, and a succession of higher-order infinitesimals compounds the incoherence, since 'to conceive a Part of such infinitely small Quantity, that shall still be infinitely less than it, and consequently though multiply'd infinitely shall never equal the minutest finite Quantity, is, I suspect, and infinite Difficulty to any Man whatsoever' (Section 5).

These metaphysical objections are not, by themselves, completely decisive, since even in Berkeley's day there were numerous examples of 'incomprehensible' magnitudes that had later become accepted (negative, irrational, and complex numbers being the most salient examples). But the case is made significantly stronger when he adds logical objections, purporting to show that the calculus contains 'Emptiness, Darkness, and Confusion; nay, If I mistake not, direct Impossibilities and Contradictions' (Section 8). If the principles and demonstrations of the calculus are logically flawed by containing incoherent or contradictory assumptions, then the rigor of its procedures is seriously compromised.

Berkeley's logical critique begins by evaluating two proofs of fundamental results in the calculus. The first is Newton's proof of the rule for determining the fluxion of a product as set out in Book II, Section 2, Lemma 2 of the *Principia* (§5, Table 3). Newton treats the product of two flowing quantities as a rectangle, whose sides are A and B. The respective moments of these flowing quantities are designated a and b. As Newton explains, flowing quantities are 'indeterminate and variable' and their moments are the instantaneous increments or decrements. To calculate the moment of the product AB he first considers the case where each flowing quantity lacks one-half of its moment. The resulting rectangle has the area expressed by $(A - 1/2a)(B - 1/2b)$, which expands to become $AB - 1/2Ab - 1/2aB + 1/4ab$. He next takes the rectangle formed after the flowing quantities have gained the remaining halves of their moments, namely: $(A + 1/2a)(B + 1/2b)$. Expanded, this yields

$$AB + 1/2Ab + 1/2aB + 1/4ab. \tag{1}$$

Subtracting the first product from the second yields the moment of the rectangle AB, namely $Ab + Ba$.

Berkeley dismisses this argument as a sham intended to mask the use of infinitesimal magnitudes. As he notes, the definition of a moment as an increment requires that the

'direct and true' method of computing the moment of AB is to consider the difference between it and the augmented rectangle $(A+a)(B+b)$. This requires taking the difference between AB and $AB+Ab+aB+ab$. The resulting increment or moment is $Ab+Ba+ab$, which differs from Newton's result by the additional term ab. In effect, Newton takes the increment of the rectangle $(A-1/2a)(B-1/2b)$, and the obvious intention of this move is to avoid the annoying term ab. Berkeley concludes that 'though much Artifice hath been employ'd to escape or avoid the admission of Quantities infinitely small, yet it seems ineffectual' (Section 11). Taking an increment of a product less than AB is a handy way to avoid having to dismiss the term ab as infinitely less than either of the terms aB or Ab, but it cannot be made consistent with Newton's pronouncements on the nature of moments.

Thus Berkeley's objection not only points out a fundamental error in the Newtonian proof; it also shows the vanity of Newton's pretense to have based his methods exclusively on the consideration of finite magnitudes. Mathematicians of the Newtonian school often claimed that their method of fluxions was more rigorous than the Leibnizian calculus, which they accused of employing obscure and extravagant assumptions about infinitesimals. Berkeley, however, has shown that Newton's own procedures make a mockery of such pretensions to rigor.

The second part of Berkeley's logical criticism of the calculus focuses on Newton's rule for finding the fluxion of any power, as demonstrated in his treatise 'On the quadrature of curves', first published in 1704. He remarks that in view of the lamentable state of the proof in the *Principia* Newton must have suffered 'some inward Scruple or Consciousness of defect in the foregoing Demonstration' and therefore resolved 'to demonstrate the same in a manner independent of the foregoing Demonstration' (Section 12). Berkeley prefaces his criticism with a lemma that forbids the use of contradictory premises in a demonstration. As he phrases it (Section 12):

> If with a View to demonstrate any Proposition, a certain Point is supposed, by virtue of which certain other Points are attained; and such supposed Point be it self afterwards destroyed or rejected by a contrary Supposition; in that case, all the other Points, attained thereby and consequent thereupon, must also be destroyed and rejected, so as from thence forward to be no more supposed or applied in the Demonstration.

He characterizes this principle as 'so plain as to need no Proof' and proceeds to show that Newton's proof procedure gives the strong appearance of violating it.

Newton's proof in the 'Quadrature of curves' assumes a flowing quantity x and shows how to find the fluxion of the power x^n. As x flows or increases to $(x+o)$ the power x^n becomes $(x+o)^n$, which by binomial expansion becomes

$$x^n + nox^{(n-1)} + \frac{n(n-1)}{2}o^2x^{(n-2)} + \cdots. \tag{2}$$

As a consequence, the increment of the flowing quantity x and that of $(x+o)^n$ stand in the ratio

$$o : nox^{(n-1)} + \frac{n(n-1)}{2}o^2x^{(n-2)} + \cdots. \tag{3}$$

Dividing both terms in the ratio by o yields the ratio

$$1 : nx^{(n-1)} + \frac{n(n-1)}{2} ox^{(n-2)} + \cdots. \tag{4}$$

Newton defines the fluxion of x^n as the ultimate ratio between 'evancescent' or vanishing increments, *i.e*. the ratio that holds as the increments are diminished to nothing. By letting the increment o vanish and discarding terms that contain it, he obtains $1 : nx^{(n-1)}$ as the ratio of the flowing quantity x to the power x^n. In other words, the fluxion of x^n is $nx^{(n-1)}$.

Berkeley objects that inconsistent assumptions have been used concerning the quantity o. In the transition from equation (3) to (4) it is assumed that o is positive in order to carry out the division; but terms containing o can only dismissed from (4) to get the desired result if it is assumed that o is zero. Commenting on the transition from (4) to the final result Berkeley remarks (Section 12):

> Hitherto I have supposed that x flows, that x hath a real Increment, that o is something. And I have proceeded all along on that Supposition, without which I should not have been able to have made so much as one single Step. From that Supposition it is that I get at the Increment of x^n, that I am able to compare it with the Increment of x, and that I find the Proportion between the two Increments. I now beg leave to make a new Supposition contrary to the first, *i.e.* I will suppose that there is no Increment of x, or that o is nothing; which second Supposition destroys my first, and is inconsistent with it. I do nevertheless beg leave to retain $nx^{(n-1)}$, which is an Expression obtained in virtue of my first Supposition, and which could not be obtained without it: All which seems a most inconsistent way of arguing, and such as would not be allowed of in Divinity.

The justice of these charges was a matter of intense debate in Berkeley's day, but there is no question that his arguments show that the Newtonian calculus of fluxions is *prima facie* unrigorous.

Aside from calling the rigor and coherence of the Newtonian fluxional calculus into question, Berkeley argued that there was no useful distinction between it and the Leibnizian *calculus differentialis*. This charge had a significant *ad hominem* effect in the context of Newtonians' claims for the superior rigor of their procedures in comparison with those of the Leibnizian school. After remarking that Newton's method is 'in effect the same with that used in the *calculus differentialis*' because it requires a 'marvellous sharpness of Discernment, to be able to distinguish between evanescent Increments and infinitesimal Differences' (Section 17), Berkeley echoes the Newtonian complaints against Leibnizian infinitesimal differences by arguing that the Leibnizians make 'no manner of scruple, first to supposed, and secondly to reject Quantities infinitely small: with what clearness in the Apprehension and justness in the reasoning, any thinking man, who is not prejudiced in favour of these things, may easily discern' (Section 18). The result is that Newtonian criticisms of the Leibnizian calculus are turned against the calculus of fluxions itself, and the foundations of the calculus are rendered obscure and burdened with apparent self-contradiction.

4 RESPONSES TO BERKELEY

The publication of *The analyst* touched off an intense controversy in the British mathematical community. In the decade after 1734 numerous writings appeared in reply to Berkeley, all offering interpretations of the calculus designed to overcome his objections. There was nevertheless considerable disagreement over how to interpret the foundational concepts of the calculus so as to deflect Berkeley's criticisms, and some who defended the rigor of the calculus became involved in disputes among themselves over how best to proceed. Surveying the British mathematical landscape in the aftermath of *The analyst*, Florian Cajori aptly characterized Berkeley's arguments as 'so many bombs thrown into the mathematical camp' [1919, 57].

Two noteworthy respondents to The analyst were James Jurin and Benjamin Robins, whose differing approaches led them into a long-running controversy over the nature of evanescent magnitudes and limiting processes. Jurin's *Geometry no friend to infidelity* (1734) attempted to defend Newton's reasoning at every turn, even to the point of insisting that a ratio of evanescent increments could subsist even as the quantities forming the ratio vanish. On Jurin's analysis, there is no inconsistency in dividing by an the increment o to simplify a ratio and then dismissing any remaining o-terms as 'vanished'.

Robins's *Discourse concerning the nature and certainty of Sir Isaac Newton's methods of fluxions and of prime and ultimate ratios* also appeared in 1734 and developed a very different defense of Newton's methods. He provided complex exhaustion proofs in the style of Archimedes to justify the basic techniques of the calculus, considering only finite differences between finite magnitudes and using arguments by *reductio ad absurdum* to establish central results instead of appealing to evanescent increments. On his account of the matter, Newton's concise way of expressing himself had given rise to several confusions and misinterpretations on which Berkeley's critique was based. Robins published a review of the *Analyst* controversy, in which he argued (*contra* Jurin) that ultimate ratios must be taken as limits of sequences of finite ratios, not ratios of vanishing quantities. Jurin replied, and the resulting controversy dragged on for a number of years without definitive resolution.

Aside from Jurin and Robins, the contributions of Colin Maclaurin and Roger Paman deserve mention. Maclaurin's *Treatise of fluxions* (1742) was written in response to *The analyst* and undertook the rigorization of the calculus of fluxions by basing its fundamental definitions on Newton's kinematic theory of magnitudes and then deriving the main results with exhaustion proofs in the style of Archimedes (§10). Thus, Maclaurin conceives of curves as generated by the motion of points and defines the fluxion of a curve as the (directed) velocity with which it is generated; in the case of uniform motion this is the ratio of the distance the point travels to the elapsed time, while for accelerated motions the instantaneous velocity becomes the distance the point would travel in a unit time were it to continue unaccelerated from that instant.

Paman's *Harmony of the ancient and modern geometry asserted* appeared in 1745. Avoiding any talk of motion or acceleration, he worked out an approach to fluxions that considers only finite differences of magnitudes and bears a surprisingly strong resemblance to the modern view of the calculus. He introduced the terms *minimaius* and *maximinus*, defined over a class of quantities in a way very similar to the contemporary definitions of

least upper bound and greatest lower bound. He then put them to much the same use as these contemporary concepts, defining the fluxion of a curve as the 'first minimajus' or 'last maximinus' of a sequence of quantities approximating the tangent. Although Paman's work was little read, both he and Maclaurin provided responses to *The analyst* that show Berkeley was right about fundamental obscurities in the calculus while also indicating how to overcome them.

BIBLIOGRAPHY

Blay, M. 1986. 'Deux moments de la critique du calcul infinitésimal', *Revue d'histoire des sciences, 39*, 223–253.

Breidert, W. 1986. 'Berkeleys Kritik an der Infinitesimalrechnung', in *300 Jahre 'Nova methodus' von G.W. Leibniz* (ed. A. Heinekamp), Stuttgart: Steiner (*Studia Leibnitiana*, Sonderheft 14), 185–191.

Breidert, W. 1989. *George Berkeley, 1685–1753*, Basel, Boston, and Berlin: Birkhäuser (*Vita mathematica*, vol. 4).

Cajori, F. 1919. *A history of the conceptions of limits and fluxions in Great Britain from Newton to Woodhouse*, Chicago: Open Court.

Cantor, G. 1984. 'Berkeley's *Analyst* revisited', *Isis*, *75*, 668–683.

Grattan-Guinness, I. 1969. 'Berkeley's criticism of the calculus as a study in the theory of limits', *Janus*, *56*, 215–227. [Printing correction, 57 (1970: publ. 1971), 80.]

Hughes, M. 1992. 'Newton, Hermes, and Berkeley', *British journal for the philosophy of science*, *43*, 1–19.

Jesseph, D.M. 1993. *Berkeley's philosophy of mathematics*, Chicago: University of Chicago Press.

Sherry, D. 1987. 'The wake of Berkeley's *analyst*: rigor mathematicae?', *Studies in history and philosophy of science*, *24*, 455–480.

Stock, J. 1776. *An account of the life of George Berkeley*, London: J. Murray.

Wisdom, J.O. 1939. 'The *Analyst* controversy: Berkeley's influence on the development of mathematics', *Hermathena*, *54*, 3–29.

Wisdom, J.O. 1942. 'The *Analyst* controversy: Berkeley as a mathematician', *Hermathena*, *59*, 111–128.

CHAPTER 9

DANIEL BERNOULLI, *HYDRODYNAMICA* (1738)

G.K. Mikhailov

Besides introducing the first hydraulic theory of the fluid flow, this book is the most remarkable general work in theoretical and applied mechanics written in the pre-Lagrangean period of the 18th century, based on a deep physical understanding of mechanical phenomena and presenting many new ideas for the following scientific progress.

First publication. Hydrodynamica, sive de viribus et motibus fluidorum commentarii. Opus academicum ab auctore, dum Petropoli ageret, congestum, Argentorati [= Strasbourg]: J.R. Dulsecker, printed by J.H. Decker, 'Basiliensis', 1738. [vi] + 304 pages + 12 tables with 86 figures.

New edition. In D. Bernoulli *Werke*, vol. 5 (ed. G.K. Mikhailov), Basel: Birkhäuser, 2002.

Russian translation. *Gidrodinamika, ili zapiski o silakh i dvizheniyakh zhidkostei* (trans. V.S. Gokhman, with commentaries by A.I. Nekrasov and K.K. Baumgart), [Leningrad]: USSR Academy of Sciences, 1959. [Also many historical comments (including extracts from the draft manuscript of *Hydrodynamica*) and survey of Bernoulli's life and work by V.I. Smirnov.]

German translation. *Hydrodynamik oder Kommentare über die Kräfte und Bewegungen der Flüssigkeiten* (trans. K. Flierl), Munich: *Veröffentlichungen des Forschungsinstituts des Deutschen Museums für die Geschichte der Naturwissenschaften und der Technik*, series C, *Quellentexte und Übersetzungen*, no. 1a, 1965. [Editor's remarks, with mathematical misprints, in no. 1b.]

English translation. *Hydrodynamics* (trans. T. Carmody and H. Kobus), in D. Bernoulli *Hydrodynamics* and J. Bernoulli *Hydraulics*, New York: Dover, 1968, xvii–xxii, 1–342. [Translation not always accurate [Binnie and Easterling, 1969].]

Manuscript. Draft in the Petersburg Archives of the Russian Academy of Sciences; to appear in Bernoulli *Werke*, vol. 4.

Related articles: Newton (§5), d'Alembert (§11), Lagrange on mechanics (§16).

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

1 DANIEL BERNOULLI'S LIFE AND WORK

Daniel Bernoulli (1700–1782) belonged to the well-known family of Swiss mathematicians; he was the second son of Johann I Bernoulli (1667–1748). Born in Groningen, he came as a child, with his father's family, to Basel where he studied at the University. During almost two years the young Bernoulli tried to practice as a physician in Italy and was then invited, as a member of the illustrious family solely, to the Saint Petersburg Academy of Sciences that was organised in 1725. During over seven years of work in Russia he developed as a great scientist in the field of pure and applied mathematics, confirming the family's fame. In 1733 he returned to Basel and lived there permanently thereafter, from 1733 as Professor of Anatomy and Botany, and from 1750 as Professor of Physics. He was generally recognised, honoured by election at the most prestigious European academies of Saint Petersburg, Berlin and Paris, as well as at the Royal Society of London. However, he never accepted flattering invitations from the Berlin and Saint Petersburg sovereigns to leave Basel again.

Daniel Bernoulli published about 80 works, including 50 papers in the editions of the Petersburg Academy of Sciences and 10 prize-winning memoirs of the Paris Academy. But there was only one large treatise—his famous *Hydrodynamica*, which had a complicated and partially dramatic fate.

2 GENERAL REMARKS

The contents of the *Hydrodynamica* are summarised in Table 1; its full title may be rendered 'Hydrodynamics, or commentaries on the forces and motions of fluids'. In this book Bernoulli presented the earliest adequate theory of motion of an incompressible fluid in tubes (vessels) and fluid outflow through orifices, introducing the notion of the hydrodynamic pressure. However, the treatise is not restricted to theoretical hydraulics. In the subsequent sections, he opens up new branches of physics and mechanics. He develops the first model of the kinetic theory of gases, approaches the principle of conservation of energy, establishes a foundation for the analysis of efficiency of machines, and he develops a theory of hydroreactive (water-jet) ship propulsion, including a solution of the first problem of motion of a variable-mass system (see section 4 below).

Hydrodynamica contains many profound remarks on the physical background of a wide range of mechanical effects, and its study remains most edifying also to the modern reader. Bernoulli's treatise was to influence the entire development of mechanics and, especially, of applied mechanics, for at least a century. However, many of his advanced ideas were far ahead of his time and met an adequate understanding only later. In the 19th century, J.-V. Poncelet called Bernoulli's treatise 'the immortal *Hydrodynamica*' in 1845, and Paul Du Bois-Reymond referred to 'the enormous wealth of ideas which assures this work one of the first places in the literature of Mathematical Physics of all ages' in 1859.

Hydrodynamica is founded mainly on the *principle of conservation of 'living forces'* (that is, kinetic energy). Bernoulli preferred to use this principle not in its traditional form, received with hostility by Newtonians, but in Christiaan Huygens's formulation that

Table 1. Summary by Sections of Bernoulli's *Hydrodynamica*.

Sect.	Page	Topics
I	1	General introduction: historical survey, general principles of the theory.
II	17	Hydrostatics.
III	30	Fluid velocity at outflow from vessels.
IV	61	Duration of outflow.
V	90	Outflow from permanently full vessels.
VI	111	Fluid oscillation in tubes (vessels).
VII	124	Outflow from submerged vessels, and lost of living forces.
VIII	143	Flow through compound vessels with regard to loss of living forces.
IX	163	Hydraulic machines and their efficiency; mechanical work.
X	200	Kinetic model of air, properties and motion of gases, cannon shooting.
XI	244	Rotational fluid motion and fluid flow in moving vessels.
XII	256	Pressure of moving fluids on the tube (vessel) walls.
XIII	278	Reaction and impact of outflowing jets, propelling ships by jet ejection. [End 304.]

Bernoulli named the *principle of equality between the actual descent and potential ascent*: 'If any number of weights begin to move in any way by the force of their own gravity, the velocities of the individual weights will be everywhere such that the products of the squares of these velocities multiplied by the appropriate masses, gathered together, are proportional to the vertical height through which the centre of gravity of the composite of the bodies descends, multiplied by the masses of all of them'.

As to hydraulics proper, Bernoulli's considers only quasi-one-dimensional fluid motion, reducing any flow to this case by means of the *hypothesis of plane sections*: he does not distinguish between tubes and vessels. The principle of conservation of living forces was used for studying the fluid flow by Bernoulli and Leonhard Euler also earlier. Coincidence of their results presented independently in the Petersburg Academy of Sciences in 1727 forced Euler to change his scientific plans and to leave this field for his elder colleague [Mikhailov, 2000]. When Bernoulli developed his work, besides studying many special cases of flow, he achieved two new fundamental results. He succeeded in explaining the nature and determining the value of the hydrodynamic pressure of moving fluids on the wall of tubes and he discovered the principal role of losses of living forces in the fluid flow, especially at sudden changes of the flow cross-sections. The former gave an instrument to engineers for calculation of tube strength and the latter served, in addition, a step to the general *principle of conservation of energy*. Bernoulli concluded also the sharp discussion of many years on the impact and reaction of emitting jets, giving the final solution of the problem.

It should be recognized that the term *hydrodynamics* introduced by Daniel Bernoulli for the whole theory of fluid motion is now applied only for the general fluid dynamics, but is not used for the quasi-one-dimensional (hydraulic) theory of fluid flow that was in fact developed by Bernoulli.

3 THE BERNOULLI EQUATION IN THE *HYDRODYNAMICA*

Bernoulli's original derivation of the formula, now called the Bernoulli equation, is difficult to follow for the modern reader and deserves therefore detailed consideration. Drawing upon Figure 1, he formulates the problem in the following manner (Section XII, art. 5): 'Let a very wide vessel *ACEB*, which is to be kept constantly full of water, be fitted with a horizontal cylindrical tube *ED*; and at the extremity of the tube let there be an orifice o emitting water at a uniform velocity; the pressure of the water against the walls of the tube *ED* is sought'.

In Bernoulli's notation (see section 6 below), the velocity v_2 of the outflow of fluid (water) from the opening equals $\sqrt{a}$, where a is the height of the water level in the vessel above the orifice. If the ratio of the cross-section of the tube to that of the outlet Ω_1/Ω_2 equals n, then the velocity in the tube is $v_1 = \sqrt{a}/n$ (in modern terminology, $V_2 = \sqrt{2ga}$, where we use capital letter V for the real velocity, in order to avoid misunderstanding). In the case that the end section of the tube *FD* is completely open, the velocity of the fluid in it would, obviously, be v_2. However, the partial closing of the section *FD* by a cover with the opening o prevents free escape of fluid, causes its compression, and thereby creates a pressure in the flow and on the wall of the tube. 'Thus it is seen', writes Bernoulli, 'that the pressure on the walls is proportional to the acceleration or the increment of velocity which the water should receive if any obstacle to the motion vanished in an instant, so that it would be ejected immediately into the air'. Therefore, for determining the pressure on the wall of the tube, it is sufficient to imagine an instantaneous rupture of the tube wall at the section under consideration and to determine the initial acceleration of the fluid after the rupture. Bernoulli assumes that at a certain instant the tube is ruptured at the section *cd* (located at the definite distance c from the entrance section *EG*), and calculates the corresponding acceleration of the fluid motion at this section. He employs for this purpose his *principle of equality between the actual descent and potential ascent*.

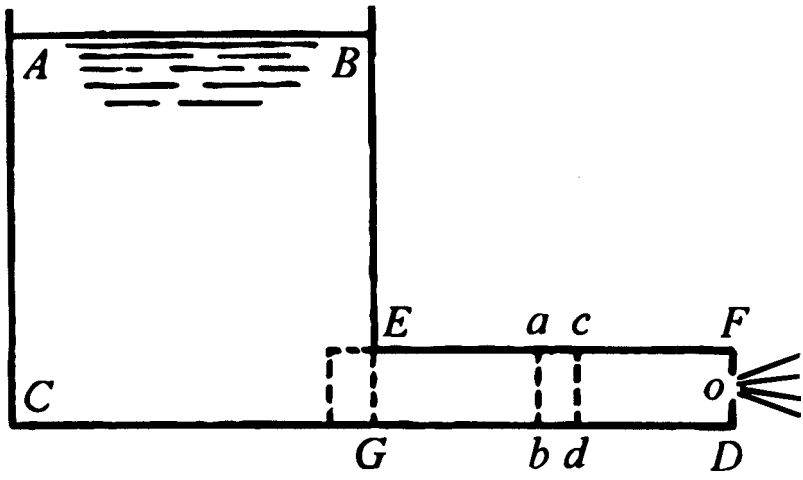


Figure 1. Bernoulli's diagram of water-flow.

Let the velocity of the fluid in the tube Ed be v $(= V/\sqrt{2g})$, and denote by *acdb* the cylindrical drop (of length $ac = dx$ and volume $\Omega_1\,dx$) passing through the section of the rupture during the elementary time interval dt. During the same time interval, a volume of fluid equal to that drop will enter the tube through its initial section EG. Then the increase in living forces over the time interval dt comprises two parts: 1) the fluid volume $\Omega_1\,dx$ entering the tube through the section EG from the wide vessel *ACEB* acquires the living force $\gamma\Omega_1\,dx v^2$ (that is, $1/2\rho\Omega_1\,dx V^2$), where ρ and γ are its density and specific weight, which Bernoulli does not introduce explicitly into the calculation; and 2) the mass of fluid in the tube Ed acquires the additional living force $2\gamma\Omega_1 cv\,dv$ (that is, $\gamma\Omega_1 cV\,dV$). Thus the entire increase of living forces during the time interval dt is $\gamma\Omega_1(v^2\,dx + 2cv\,dv)$. The *actual descent* during the same time corresponds to a descent of the volume $\Omega_1\,dx$ from the level of the free fluid surface in the vessel to the level of the tube (at height a) and amounts to $\gamma a\Omega_1\,dx$. Setting the *potential ascent* equal to the *actual descent*, one obtains

$$\gamma\Omega_1(v^2\,dx + 2cv\,dv) = \gamma a\Omega_1\,dx, \quad \text{or} \quad v\frac{dv}{dx} = \frac{a - v^2}{2c}. \tag{1}$$

However, for any motion, the acting force (pressure) is proportional to the ratio of the velocity increase over the time element. Thus, in the case under consideration, the pressure p_1 in the tube is proportional to the ratio of dv to $dt = dx/v$, that is,

$$p_1 = \alpha v\frac{dv}{dx}, \tag{2}$$

where α is a certain constant coefficient. According to the preceding equation,

$$p_1 = \alpha v\frac{dv}{dx} = \alpha\frac{a - v^2}{2c}. \tag{3}$$

But at the initial instant the velocity v in the tube is $\sqrt{a}/n$, so that

$$\frac{a - v^2}{2c} = \frac{n^2 - 1}{2n^2 c}a \quad \text{and} \quad p_1 = \alpha\frac{n^2 - 1}{2n^2 c}a. \tag{4}$$

There remains to find the value of the coefficient α. Since Bernoulli assumes that this coefficient does not depend on the geometrical parameters, it is sufficient for him to study the simplest case of the infinitely small hole o, when the fluid in the tube is virtually motionless and the pressure in the system 'vessel + tube' is hydrostatically distributed. Under this condition, the pressure in the tube is determined by the height of the column of fluid a (i.e., $p_1 = \gamma a$), and $n \to \infty$, whence $\alpha = 2c\gamma$. Consequently, Bernoulli obtains for the pressure in the tube the expression

$$p_1 = \gamma\frac{n^2 - 1}{n^2}a. \tag{5}$$

(In the original text of the *Hydrodynamica* this formula contains neither the specific weight γ, nor even the letter p denoting the pressure.) This is the form in which the *Hydrodynamica* presented, for the first time, the famous Bernoulli equation for the case of steady flow

of an incompressible fluid. In order to bring this expression closer to that which now bears the same name, it can be rewritten taking into consideration that $n = \Omega_1/\Omega_2$, the velocity of the fluid in the tube $V_1 = \sqrt{2ga}/n$, the outflow velocity $V_2 = \sqrt{2ga}$, $V_1\Omega_1 = V_2\Omega_2$, and the pressure p_2 at the orifice vanishes. Thus

$$\frac{p_1}{\gamma} + \frac{V_1^2}{2g} = \frac{p_2}{\gamma} + \frac{V_2^2}{2g}. \tag{6}$$

Bernoulli did not pay attention to the fact that his reasoning concerned only the case of steady flow. A general solution for unsteady quasi-one-dimensional flow is given by Daniel's father Johann Bernoulli in the second part of his *Hydraulica*, and the generalized Bernoulli equation, usually known as the Lagrange–Cauchy equation, was obtained later in [Euler, 1757].

4 PROPULSION OF SHIPS BY JET EJECTION, AND THE DYNAMICS OF SYSTEMS WITH VARIABLE MASS

In Section XIII, Bernoulli advances the idea of exploiting the reaction of an emitting water jet for the propulsion of ships, and studies the motion of a hydroreactive craft. This part of the *Hydrodynamica* is of interest for two reasons. On the one hand, the subsequent development of water-jet ship propulsion is generally linked to Bernoulli's suggestion; on the other hand, Bernoulli was the first to consider here a problem of motion of a body with variable mass and devise a simple solution based on the *law of conservation of momentum*.

The dynamics of systems with variable mass—of which the motion of a hydroreactive craft presents a particular case—is an elementary part of general dynamics. Nevertheless, the basic equations of this branch of dynamics have been rediscovered frequently and have evoked wide discussion over two centuries. In the 20th century, the dynamics of systems with variable mass has become especially important in connection with the rapid development of rockets and space technology. But, despite their apparent simplicity, the equations of motion of a particle with variable mass have sometimes defied comprehension even in the middle of the 20th century [Mikhailov, 1976].

Bernoulli's idea of water-jet ship propulsion and its elementary theory is explained in the last part of Section XIII. 'It entered my mind at one time', Bernoulli begins, 'that these things which I had pondered about the repelling force of fluids while they are ejected [...] can be applied usefully to instituting a new method of seafaring. For I do not see what should prevent very large ships from being moved without sails and oars by the method that water is continually elevated to a height and then flows out through orifices in the lowest part of the ship, contriving that the direction of the water flowing out faces towards the stern'.

Bernoulli now goes on to calculate the efficiency of this mode of propulsion, starting from the observation that 'a ship is retarded continuously by the water drawn in on account of its inertia, when the same velocity is communicated to it at which the ship is borne, and while it is communicated, the ship is forced backwards by the reaction of the water, but at the same time it is pressed forward by the outflow of the same'. Thus he clearly poses the particular problem of the motion of an object of variable mass (or of variable

composition of its constant total mass). He has already determined the reaction force of the outflowing jet, and it remains for him to analyse 'the resistance of the ship due to continually receiving into it water from the quiescent basin on which it is moving'. With the help of the *law of conservation of momentum* and by equating the inflow of water into the ship to the discharge of the ejecting stream, Bernoulli easily finds the 'resistance' to be (in modern notation)

$$R = V_2 \rho \Omega V_1, \tag{7}$$

where V_2 is the velocity of the ship, V_1 the relative outflow velocity of water from the opening in the stern, and Ω is the area of this opening. Since the reaction force of the outflowing water is $P = \rho \Omega V_1^2$, he finds for the total force propelling the ship

$$F = P - R = \rho \Omega V_1^2 - V_2 \rho \Omega V_1 = \rho \Omega V_1 (V_1 - V_2). \tag{8}$$

As has been emphasized, this calculation of the propulsive force of a hydroreactive craft is the first analysis in mechanics of the motion of an object of variable mass.

The mechanical foundations of the motion of hydroreactive craft were mentioned publicly perhaps only once later on in the 18th century. Discussing Bernoulli's new idea of ship propulsion, Benjamin Franklin emphasized that it was necessary to take into account the resistance of the 'inertia force' of the water reaching the ship, which would decrease the 'moving power'. Interest in the hydro-jet propulsion of ships began to revive in the middle of the 19th century, but then usually only Euler's definition of the reactive force of a fluid flowing out from a nozzle was cited.

5 SOME OTHER TOPICS

5.1 The principle of conservation of energy. The crucial point of Daniel Bernoulli's hydraulics is that he took into account possible losses of a part of the living forces during the fluid flow 'because it often occurs that the motion goes over to other matter'. His insight into the actual and potential living forces is closely linked to the notion of energy introduced in the 19th century into thermodynamics and mechanics. A certain general conservation principle was actually used by Bernoulli, as he assumed the possibility of a transition of *living forces* (mechanical energy) from the macromotion of fluid not only to some kind of 'useless' internal mechanical motions of the fluid particles, but also to elementary particles moving due to the heat, as well as to various kinds of fine materials (*materia subtilis, insensibilis*). Both the *principle of conservation of energy* and the *principle of conservation of mass* are practically applied by him especially when dealing with physico-chemical reactions.

5.2 The kinetic theory of gases. Bernoulli proposes a kinetic model of air, consisting of a number of very small (but finite) spherical particles moving in straight lines at very high velocities. He determines the air pressure on the walls of a vessel in terms of the collisions of the particles with the walls, that is, their frequencies and their impulses. Assuming that heat increases the velocity V of particles, he shows that the elasticity (pressure) of the air is proportional to V^2 and proposes to measure the air temperature ('*aëris calor*') by this

pressure at constant density, which makes the temperature proportional to V^2, anticipating in this respect the Kelvin scale of absolute temperature. Taking into account the finite size of the air particles, Bernoulli obtains also a generalization of the Boyle–Mariotte law in the spirit of the Van der Waals equation. Unfortunately, Bernoulli's kinetic theory was disregarded over more than a century. A similar model was again proposed in the 1850s, and only in the 1870s the significance of Bernoulli's theory as a precursor of the new kinetic theories was understood and generally acknowledged in the context of mechanistic heat theories.

5.3 Barometrical studies. Having elucidated the principles of his kinetic approach, Bernoulli discusses the barometric and thermal characteristics of the atmosphere and tries to come upon a universal relationship between the barometric pressure and the height of a location above sea level. As a particular problem, the refraction of light in the atmosphere, due to the change of air density at various heights, is investigated.

5.4 The outflow of gases from vessels. The *Hydrodynamica* includes the first attempts to study the flow of gases. Of course, Bernoulli is limited by analysing isothermal gas flows only, as it was not known at that time whether the air temperature varies during compression (the notion of adiabatic flow was introduced by Laplace much later). The solutions obtained are approximate: Bernoulli transfers his hydraulic method to air, considering the flow as a sequence of steady states: even Louis Navier analysed isothermal gas flow in the 1820s under the same assumptions.

5.5 The work of compressed air and gunpowder gases. Bernoulli discusses the use of compressed air as a source of movement and evaluates the force of gunpowder gases according to the work they can perform. He introduces the notion of the potential living force of an elastic fluid at rest, 'wherein nothing else is understood by this than the *potential ascent* which an elastic body can communicate to other bodies before it will have lost all its elastic force'. By exploiting this notion, he explains the possibility of driving machines by means of heated (or cooled) air and discusses the force of ignited gunpowder for projecting missiles. He notes 'that the effect of [...] one pound of ignited gunpowder for elevating weights can be greater than that which one hundred very robust men can accomplish by continuous labour within one day's span'.

5.6 The efficiency of machines. Bernoulli presents some general principles for an evaluation of the performance of machines. He limits his analysis of hydraulic machines to their steady operation, but introduces the fundamental notions of *work* ('*potentia absoluta*') and *efficiency*, which are very important for the general theory of machines. In particular, he estimates the theoretical efficiency of elementary hydraulic machines, taking into account the loss of work connected with the creation of redundant flow velocities and with water leakage, the loss of kinetic energy at outflow through orifices, and mentioning the loss due to mechanical friction. From this point of view, Bernoulli analyses construction of certain types of pumps, giving recommendations for their improvement. The next large step in the development of the theory of machines occurs in the first half of the next century and is due to the French school of applied mechanics, with which the broad use of the notion of *work* is traditionally associated.

5.7 Varia. Discussing problems of fluid flow, Bernoulli touches upon phenomena such as the water hammer, turbulent character of fluid motion, and cavitation. The *Hydrodynamica* includes also many pure mathematical problems connected with solving differential equations, integrations and series.

6 SOME GENERAL REMARKS ON THE STYLE OF THE *HYDRODYNAMICA*

In reading Daniel Bernoulli's work, it is necessary to remember that there did not yet exist any theory of dimensions of physical quantities in the 18th century, and scientists used different systems of physical units. Bernoulli usually employed a system of physical units which was based not on three basic units, as in modern times, but only on two, namely on the units of length L and of force (weight) F. Reduction of the number of basic physical units is equivalent to the introduction of an additional dimensionless quantity. In this case, this additional dimensionless quantity is the gravitational acceleration. As a consequence, the dimension of mass $[M]$ does not differ from that of force $[F]$. Correspondingly, the dimensions of density ρ and specific weight γ likewise coincide: $[\rho] = [\gamma] = [FL^{-3}]$. It also follows directly from the definition of acceleration that the dimension of time $[T]$ is $[L^{1/2}]$. Since velocity is defined as the ratio of a distance to time, its dimension $[V]$ is likewise $[L^{1/2}]$.

In works of the first half of the 18th century, velocity is evaluated, as a rule, by a linear quantity, namely velocity head, and is defined as the square root of the velocity head $\sqrt{H}$ or $\sqrt{2H}$. Bernoulli employs the first option ($v = \sqrt{H}$) throughout most of his *Hydrodynamica*. Therefore, when transforming his formulae written for the given determination of velocity into their modern form, one must replace in them all quantities according to the scheme (we use capital letters for the up-to-date defined physical values)

$$\text{mass} \to Mg, \quad \text{velocity} \to \frac{V}{\sqrt{2g}}, \quad \text{time} \to T\sqrt{2g}. \tag{9}$$

Curiously, in this physical system of units, the basic equation of dynamics $F = MdV/dT$ becomes $f = 2m\,dv/dt$, with the coefficient 2, which is strange to the modern reader. The living force $\frac{1}{2}MV^2$ is represented in this system as the product of weight or mass by the square of velocity v^2 $(= H)$ without the factor $1/2$. In the second case ($v = \sqrt{2H}$), the basic law of dynamics, most deceivingly, retains its familiar form $f = m\,dv/dt$, although, as before, velocity and time have here the dimension $[L^{1/2}]$; living force as well is represented in this case in a form familiar to the modern reader, that is, by $\frac{1}{2}mv^2$. Such a variant of the system of physical units is employed in the *Hydrodynamica* in some parts of Sections IX and X, and in the entire Section XIII. Of course, in purely descriptive examples, Bernoulli evaluates velocity in the familiar way as the distance covered by the object in unit time.

It should be noted that Bernoulli did not prepare his book carefully enough for the printer. It is only in this way that we can explain, for example, that there remain in the book some obsolete, uncorrected cross-references to its separate sections, and Section XIII relies on a velocity measure which differs from the one used in all preceding Sections. Correcting in Section XIII his previous false estimate of the jet impact, Bernoulli did not introduce

any changes into the text of Section IX, where he had used the erroneous estimate, and restricted himself to the corresponding statement in art. 15 of Section XIII.

7 EARLY PRAISE AND TROUBLES

Daniel Bernoulli wrote the first version of his *Hydrodynamica* in Petersburg at the beginning of the 1730s. On his departure from Petersburg in summer 1733, he left behind a copy of the draft manuscript. The first detailed information on the forthcoming release of the *Hydrodynamica* appears in the September 1734 issue of the journal *Mercure Suisse*. In December 1734 Bernoulli writes to Euler with satisfaction: 'My *Hydrodynamica* is now really being printed by Mr. Dulsecker, and he gives me, besides 30 copies, even 100 thalers of royalty'. However, in March 1736, Bernoulli writes in desperation: 'My *Hydrodynamica* has fallen into the hands of a man who deals very badly with it: I doubt that it will ever see the light of day'.

The actual printing of the book seems to have begun only in 1737. It appeared at the end of April or the beginning of May 1738. In May, Bernoulli sends first copies of the treatise to Petersburg and asks Euler to inform on his remarks. However, on its way to Petersburg the book parcel went astray. Expecting Euler to have received a copy of the *Hydrodynamica*, Bernoulli asks him repeatedly for advice with a view to a second edition regarding changes and corrections and calls his attention especially to the last five sections of the treatise which he believes 'can contribute not in a small way to the perfection of physics, mechanics, and so on'.

Bernoulli's treatise reached Petersburg and Euler only in spring 1739. Euler reports to Daniel that he has, finally, become acquainted with his book:

> Meanwhile I have read through your incomparable Treatise with full attention and have drawn immense gain from it. I congratulate you, Sir, from all my heart on the felicitous execution of such a difficult and obscure topic, as well as on the immortal fame thus gained. The entire execution of the project deserves all conceivable attention, and all the more so as it is not accessible to rigorous mathematics, but demands the help of several important physical principles, which you have known to employ to indescribable advantage [. . .] In the case of a new edition of this Treatise, I would especially humbly advise you to set out most topics in some more detail, partly for the reader's convenience, but mainly to ensure that the great usefulness to be gained from many investigations is highlighted more prominently: In fact, I have encountered therein so many different, important and completely new topics that most of them would indeed deserve to be treated in separate publications.

By contrast, having become acquainted with Daniel Bernoulli's *Hydrodynamica*, his father Johann at once began to prepare surreptitiously his own version of the science of hydraulics. The first mention of this work known to us occurs in the letter of Johann Bernoulli to Euler of 11 October 1738, and Bernoulli sent the first, preliminary part of his *Hydraulica* to Petersburg on 7 March 1739, withholding it consciously from his son. Only someone with such extraordinary insight as Euler could immediately recognize the jewel in Johann's

Hydraulica after looking through its first part. Moreover, Euler virtually drafted the basis on which the second, most important part of the *Hydraulica* should be rested in his reply letter to the elder Bernoulli. The latter could prepare this part and sent it to Euler only in August 1740. Not waiting for the publication of his *Hydraulica* in Petersburg, Johann Bernoulli included it into his *Opera omnia* published at the beginning of 1743 (with the year 1742 on the title page), supplying it here with an extra subtitle: *now for the first time disclosed and directly shown from purely mechanical foundations, 1732.*

After Johann Bernoulli's *Opera omnia* have appeared, Daniel complains to Euler about what seemed to him to be an extreme injustice. On 4 September 1743 he writes [Fuss, 1843, 530–532]:

> Of my entire *Hydrodynamica*, not one iota of which do in fact I owe to my father, I am all at once robbed completely and lose thus in one moment the fruits of the work of ten years. All propositions are taken from my *Hydrodynamica*, and then my father calls his writings *Hydraulica, now for the first time disclosed, 1732*, since my *Hydrodynamica* was printed only in 1738. All this my father has taken over from me, except that he has thought up another method of determining the increment of velocity, which discovery occupies some few pages. What my father does not fully ascribe to himself, he contemns, and finally, to top my misfortune, he inserts your letter, in which my discoveries (of which I am completely the first and indeed the only author, and which I consider to have exhausted completely) in some measure belittle [. . .] At first it was sheer unbearable to me; but finally I have taken it all with some resignation; but also I have conceived a disgust and revulsion for my former studies, so that I had liefer learned cobbling than mathematics.

The truth seems to be that Johann Bernoulli consciously falsified the dating of his undoubtedly very interesting work on fluid dynamics that was definitely written between 1738 and 1740, stimulated by his son's *Hydrodynamica*, and could not even have been conceived as far back as 1732.

Daniel Bernoulli discussed the preparation of a second edition of the *Hydrodynamica* already in summer 1738 and at the end of 1740 he negotiated for a new edition of his treatise in French. However, neither a new Latin edition, nor the French translation of the *Hydrodynamica*, ever appeared. What is more, Daniel abandoned his further research in fluid dynamics, possibly changing thereby the subsequent development of this field of science. The reason was doubtless the heavy psychological trauma inflicted on him by his father. Thus, from many points of view, the fate of the *Hydrodynamica* was dramatic.

BIBLIOGRAPHY

d'Alembert, J. 1744. *Traité de l'équilibre et du mouvement des fluides, pour servir de suite au Traité de dynamique*, Paris: David l'ainé. [2nd ed. 1770.]

Bernoulli, J. 1742. 'Hydraulica nunc primum detecta ac demonstrata directe ex fundamentis pure mechanicis. Anno 1732', in his *Opera omnia*, vol. 4, Lausanne et Genève: Bousquet, 387–488. [Repr. Hildesheim: Olms, 1968.]

Berthold, G. 1876. 'Daniel Bernoulli's Gastheorie, eine historische Notiz', *Annalen der Physik und Chemie, (6) 9*, 659–661.

Binnie, A.M. and Easterling, H.J. 1969. Review of *Hydrodynamics* by D. Bernoulli and *Hydraulics* by J. Bernoulli, *Journal of fluid mechanics*, *38*, 855–856.

Euler, L. 1757. 'Principes généraux du mouvement des fluides' and 'Continuation', *Mémoires de l'Académie des Sciences et Belles-Lettres de Berlin, 11* (1755), 274–315, 316–361. [Repr. in *Opera omnia*, ser. 2, vol. 12, 54–91, 92–132.]

Fuss, P.-H. (ed.) 1843. *Correspondance mathématique et physique de quelques célèbres géomètres du XVIIIème siècle*, vol. 2, Saint Petersburg: [Academy of Sciences].

Mikhailov, G.K. 1976. 'On the history of variable-mass system dynamics', *Mechanics of solids (MTT)*, 10 (1975), no. 5, 32–40.

Mikhailov, G.K. 1996. 'Early studies on the outflow of water from vessels and Daniel Bernoulli's *Exercitationes quaedam mathematicae*', in D. Bernoulli *Werke*, vol. 1, Basel: Birkhäuser, 199–255.

Mikhailov, G.K. 2000. 'The origins of hydraulics and hydrodynamics in the work of the Petersburg academicians of the 18th century', *Fluid dynamics*, *34* (1999), 787–800.

Mikhailov, G.K. 2002. 'Introduction to Daniel Bernoulli's *Hydrodynamica*', in D. Bernoulli, *Werke*, vol. 5, Basel: Birkhäuser, 17–86.

Pacey, A.J. and Fisher, S.J. 1967. 'Daniel Bernoulli and the vis viva of compressed air', *British journal for the history of science*, *3*, 388–392.

Szabó I. 1987. 'Über die sogenannte Bernoullische Gleichung der Hydromechanik; die Stromfadentheorie Daniel und Johann Bernoullis', in his *Geschichte der mechanischen Prinzipien und ihrer wichtigsten Anwendungen*, 3rd ed., Basel: Birkhäuser, 157–198. [1st ed. 1977, 2nd ed. 1979.]

Truesdell, C.A. 1954. 'Rational fluid mechanics, 1687–1765', in L. Euler, *Opera omnia*, ser. 2 vol. 12, vii–cxxv.

Truesdell, C.A. 1968. *Essays in the history of mechanics*, Berlin: Springer.

CHAPTER 10

COLIN MACLAURIN, *A TREATISE OF FLUXIONS* (1742)

Erik Sageng

MacLaurin provided a rigorous foundation for the method of fluxions based on a limit concept drawn from Archimedian classical geometry. He went on to demonstrate that the method so founded would support the entire received structure of fluxions and the calculus, and to make advances that were taken up by continental analysts.

First publication. 2 vols., Edinburgh: T.W. and T. Ruddimans, 1742. vi + 763 pages, paginated continuously, 40 pages of plates.

Second edition. *A treatise on fluxions* [. . .] *to which is prefixed an account of his life. The whole carefully corrected and revised by an eminent mathematician. Illustrated with forty-one copperplates*, 2 vols., London: William Baynes and William Davis, 1801.

Full French translation. *Traité des fluxions* (trans. Pezenas), 2 vols., Paris: Jombert, 1749.

Abridged French translation. *Abregé du calcul intégral ou méthode inverse des fluxions: où l'on explique les moyens de découvrir les intégrales par les quadratures à l'usage du Collège Royal* (trans. Ch. Le Monnier), Paris: Charles-Antoine Jombert, 1765.

Related articles: Newton (§5), Leibniz (§4), Berkeley (§8), Euler on the calculus (§14), Lagrange on the calculus (§19).

1 COLIN MACLAURIN (1698–1746)

Colin MacLaurin was born in Kilmoden, Argyll, Scotland in February 1698. He entered the University of Glasgow in 1709, where he studied under Robert Simson, known for his *Elements of Euclid* (1756) and his restorations of Apollonius. In 1713, MacLaurin defended a thesis on the power of gravity, and was awarded the degree of master of arts. In 1717 he was appointed to the chair of mathematics at Marischal College, Aberdeen. During these years he published two papers in the *Philosophical transactions* of the Royal Society;

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

on the strength of these and of the manuscript to his first major publication, the *Geometria organica*, he was admitted to membership of the Society and was introduced to Newton, then its President, to whom he dedicated the *Geometria*, and under whose imprimature it was printed in 1720.

In November of 1725 MacLaurin accepted the position of deputy and successor for James Gregory at the University of Edinburgh. He remained a popular and influential Professor of Mathematics at Edinburgh until his death in 1746, teaching pure and applied mathematics, optics, astronomy, and experimental philosophy. *A treatise of algebra*, published posthumously in 1748, is composed of teaching materials used by MacLaurin at Edinburgh, and of papers on equations with 'impossible' (i.e. complex) roots published in the *Philosophical transactions* in 1726 and 1729, in which he extended Newton's work in his *Arithmetica universalis* (1707).

Some time after Newton died in 1727, John Conduit asked MacLaurin to collaborate in the writing of his biography, but the project seems to have lost momentum with Conduitt's death in 1737. MacLaurin continued with his part of the biography, and was reportedly dictating the final chapter, which contains a proof of the afterlife, a few hours before his death in 1746. The *Account of Sir Isaac Newton's philosophical discoveries*, published posthumously in 1748, is one of the most accessible yet least trivializing contemporary popularizations of Newton's natural philosophy.

MacLaurin took a leading role in preparing the defense of Edinburgh against the highland army of Prince Charles Stewart in the Jacobite rebellion of 1745, and when the city was occupied MacLaurin felt it was safest for him to withdraw into England. He returned, after a difficult journey both ways on horseback, including a fall and exposure to unpleasant weather, with what he described as the most dangerous cold he had ever had. He apparently never entirely recovered, and died on 14 June 1746.

2 THE *TREATISE ON FLUXIONS* (1742): FOUNDATIONS

2.1 A response to Berkeley

In 1734 George Berkeley had published *The Analyst: or, a discourse addressed to an infidel mathematician* (§10). Besides objecting to particular demonstrations and procedures, Berkeley's criticism of the method of fluxions amounted to the well substantiated assertion that it was founded inescapably either on infinitesimals or on a shifting of hypotheses, both of which were logically indefensible. MacLaurin's *Treatise* was begun partly in response to these criticisms; its contents are summarised in Table 1.

In the *Treatise*, MacLaurin founded the method of fluxions on a limit concept drawn from the method of exhaustions in classical geometry, avoiding the use of infinitesimals, infinite processes, and actually infinite quantities, and avoiding any shifting of the hypothesis. He was motivated by his belief that mathematics, properly understood, is based on real, actually existent entities, which belief made it impossible for him—as for Berkeley—to accept a system based on infinitesimals, and by his ideas about the role of mathematics in religion, both directly, as the ultimate bulwark against the skeptics, and by way of natural philosophy, the ultimate purpose of which is to support natural religion. These ideas

Table 1. Contents by chapters of MacLaurin's book. The titles of chapters are quoted. Volume I ends on page 412; Book I continues in Volume II with continuous pagination.

Ch.	**Page**	**Title**
	i–vi	Preface.
	1	Introduction.
	51	The Elements of the Method of Fluxions, Demonstrated after the Manner of the Ancient Geometricians.
	51	Book I. *Of the Fluxions of Geometrical Magnitudes.*
I	51	Of the Grounds of this Method.
II	109	Of the Fluxions of plane rectilineal Figures.
III	131	Of the fluxions of plane curvilineal Figures.
IV	142	Of the Fluxions of Solids, and of third Fluxions.
V	152	Of the Fluxions of Quantities that are in a continued geometrical Progression, the first term of which is invariable.
VI	158	Of Logarithms, and the Fluxions of logarithmic Quantities.
VII	178	Of the Tangents of curve Lines.
VIII	199	Of the Fluxions of curve Surfaces.
IX	214	Of the greatest and least Ordinates, of the points of contrary Flexion and Reflexion of various kinds, and of other affections of Curves that are defined by a common or by a fluxional Equation.
X	240	Of the Asymptotes of curve Lines, the Areas bounded by them and the Curves, the solids generated by those Areas, of spiral Lines, and of the Limits of the Sums of Progressions.
XI	304	Of the Curvature of Lines, its Variation, and the different kinds of Contact, of the Curve and Circle of Curvature, the Caustics by Reflexion and Refraction, the centripetal forces, and other Problems that have a dependence upon the Curvature of Lines.
XII	413	Of the Method of Infinitesimals, of the Limits of Ratios, and of the General Theorems which are Derived from this doctrine for the Resolution of Geometrical and Philosophical Problems.
XIII	486	Wherein the nature of the lines of swiftest descent is determined in any given hypothesis of gravity, and the problems concerning isoperimetrical figures, with other of the same kind are resolved by first fluxions and the solutions verified by synthetic demonstrations.
XIV	513	Of the ellipse considered as the section of a cylinder. Of the Gravitation towards bodies, which results from the Gravitation towards their Particles. Of the Figure of the Earth, and the Variation of Gravity Towards it. Of the Ebbing and Flowing of the Sea, and other inquiries of this nature.

Table 1. (*Continued*)

Ch.	Page	Title
	575	Book II. *Of the Computations in the Method of fluxions.*
I	575	Of the Fluxions of Quantities Considered Abstractly, or as Represented by General Characters in Algebra.
II	591	Of the Notation of the Fluxions, the Rules of the Direct Method, and the Fundamental rules of the inverse method of Fluxions.
III	615	Of the analogy between circular arches and logarithms, of reducing fluents to these, or to hyperbolic and elliptic arches, or to other fluents of a more simple form; when they are not assignable in finite algebraic terms.
IV	664	Of the area when the ordinate and base are express'd by fluents; of computing the fluents from the sums of progressions, or the sums of progressions from the fluents, and other branches of this method.
V	693	Of the general rules for the resolution of Problems. [End 754.]

led MacLaurin both to emphasize the importance of sound foundations in such a vital enterprise, and to be offended by the suggestion that mathematics is dangerous to religion, or that mathematicians are liable to lead men to infidelity.

In his introduction, MacLaurin says that geometry has been justly admired for its evidence and demonstration. 'It acquired this character by the great care of the old writers, who admitted no principles but a few self-evident truths, and no demonstrations but such as were accurately deduced from them' (p. 1). Mathematicians have fallen from this ideal, notably, as Berkeley has pointed out, with the method of indivisibles and that of divisible infinitesimals. Newton had rejected infinitesimals, says MacLaurin, and developed his method in a manner agreeable to the ancients, but so briefly as to be easily misunderstood (pp. 2–3):

> When the certainty of any part of geometry is brought into question, the most effectual way to set the truth in a full light, and to prevent disputes, is to deduce it from axioms or first principles of unexceptionable evidence, by demonstrations of the strictest kind, after the manner of the ancient geometricians. This is our design in the following treatise; wherein we do not propose to alter Sir Isaac Newton's notion of a fluxion, but to explain and demonstrate his method, by deducing it at length from a few self-evident truths, in that strict manner: and, in treating of it, to abstract from all principles and postulates that may require the imagining any other quantities but such as may be easily conceived to have a real existence. We shall not consider any part of space or time as indivisible, or infinitely little; but we shall consider a point as a term or limit of a line, and a moment as a term or limit of time: Nor shall we resolve curve lines, or curvilineal spaces, into rectilineal elements of any kind [. . .]. The method of demonstration, which was invented by the author of fluxions, is accurate and elegant; but we propose to begin with one that is somewhat different; which, being less removed from that of the ancients, may make the transition to his

> method more easy to beginners [. . .] and may obviate some objections that have been made to it.

MacLaurin presents 'the method of the ancients' with numerous examples, beginning with Euclid's proof by the misnamed method of exhaustion that circles are as the squares on their diameters. This is not an argument, as its name might suggest, that the areas of circles can be exhausted by doubling the number of sides of an inscribed polygon an infinite number of times while observing that the areas of similar polygons are always as the squares on their diagonals, so that the exhausted circles—equivalent to infinite sided polygons—must also be as the squares on their diameters. Rather it is an argument that to deny the conclusion can be shown to lead, in a finite number of steps, to a contradiction. Its method of demonstration is characteristic of almost every proof in the *Treatise*: MacLaurin always clinches his argument with a double *reductio ad absurdum*, and he never uses infinite or infinitesimal quantities or infinite processes.

MacLaurin presents numerous examples of Archimedes's application of this method to quadratures and cubatures of progressively more complex solids, and in the remainder of his introduction he narrates the development of these methods up to his time. He says that geometers have extended Archimedes's methods but abandoned his foundations, adopting, as in Bonaventura Cavalieri, indivisible or infinitesimal elements assumed infinite in number. As a result the higher geometry came to appear to be full of mysteries. MacLaurin acknowledges, as did Berkeley, the effectiveness and even subtlety of the geometry of infinites, 'but geometry is best established on clear and plain principles; and these speculations are ever obnoxious to some difficulties' (p. 47).

Bonaventura Cavalieri had been sensitive to these difficulties, but he left this Gordian knot to some Alexander. Now (pp. 49–50)

> Sir Isaac Newton [has] accomplished what Cavalierius wished for, by inventing the method of fluxions, and proposing it in a way that admits of strict demonstration, which requires the supposition of no quantities such as are infinite, and easily conceived [. . .]. In it premises and conclusions are equally accurate, no quantities are rejected as infinitely small, and no part of a curve is supposed to coincide with a right line.

Although Newton's method 'admits of strict demonstration', his own demonstrations were so brief as to be frequently misunderstood, and his method has been misrepresented as the same as the method of infinitesimals. MacLaurin says that he will explain it and 'promote the design of the great inventor, by establishing the higher geometry on plain principles, perfectly consistent with each other and with those of the ancient Geometers' (p. 50).

2.2 MacLaurin's formulation

In Chapter I, 'Of the grounds of the method', MacLaurin presents his definitions of motion and velocity. A flowing quantity is called a fluent; its velocity is a fluxion, and is measured by the increment or decrement that would be generated by the motion in a given time, if the motion were continued uniformly. All quantities—distance, time, velocity, etc.—are

represented by line lengths, and the demonstrations are all based on a limit procedure in which it is shown that quantities or ratios can be made greater than, or less than, or can approach something closer than any assignable quantity. It is obvious to him that 'while a body is supposed in motion, it must be conceived to have some velocity or other at any term of the time during which it moves', and he asserts that 'we can demonstrate accurately what are the measures of this velocity at any term'. Given this understanding of velocity, he presents four axioms that he says 'are as evident as that a greater or less space is described in a given time, according as the velocity of the motion is greater or less' (pp. 53–54, 59):

> Axiom I. The space described by an accelerated motion is greater than the space which would have been described in the same time, if the motion had not been accelerated, but had continued uniform from the beginning of the time. [. . .]
> Axiom II. The space described by a motion while it is accelerated, is less than the space which is described in an equal time by the motion that is acquired by that acceleration continued uniformly.

Axioms III and IV are analogous to I and II for the case of decelerated motion.

Using these axioms, MacLaurin proves 14 theorems, all with double *reductio ad absurdum*, about relations among quantities generated given relations among generating velocities, and vice versa. The chapter culminates with Theorem XIV in which MacLaurin makes good his claim that we may accurately know the spaces that would be described by a motion in a given time, if it were continued uniformly from some term (p. 99):

A	*P*	*D*	*G*	*g*		*a*
E	*M*	*L*	*S*	*c*	*s* *x*	*e*

> Theorem XIV. The motion of the point *P* being uniform, but the motion of the point *M* continually varied, let the velocity of *P* be to the velocity of *M* at *L*, as a given line *Dg* is to *Lc*; let *Dg* be always to *Ls*, as the space *DG* described by *P* in any time, is to *LS* the space described by *M* in the same time. Then, by diminishing the spaces *DG* and *LS* continually, *cs* may become less than any assignable magnitude.

That is, the ratio of their average speeds over an interval can become closer than any assignable difference to the ratio of their instantaneous speeds, by diminishing the interval. It is in the sense of this theorem, says MacLaurin, that we should take Newton's concept of the limit of a ratio (p. 101):

> Because *cs* the difference between *Ls* and *Lc* decreases so that it may become less than any given quantity, how small so ever, when *DG* and *LS* are diminished continually; it appears that the ratio of *Dg* to *Ls* (or of *DG* to *LS*) approaches continually to the ratio of *Dg* to *Lc*, so that it may come nearer to this ratio, than the ratio of *Dg* to any assignable quantity greater or less than *Lc*. For this reason the ratio of *Lc* to *Dg* is by Sir Isaac Newton called the Limit of the variable ratio of *Ls* to *Dg* or of *LS* to *DG*.

MacLaurin observes that *Ls* consists of two parts: *Lc*, which is invariable and measures the velocity of *M* at *L*; and *cs*, which is variable and arises from the acceleration of *M* as it describes *LS*. *cs* decreases when *DG* and *LS* are diminished, and vanishes with them. When *EM* is determined from *AP* by an equation, the ratios *LS* : *DG* or *Ls* : *Dg* are reduced to a rule, and all that is required to determine *Lc* : *Dg* (the ratio of the fluxions) is to distinguish between the variable and invariable parts of *Ls*. 'It is in this concise manner [by rejecting the variable part of *Ls*], that Sir Isaac Newton most commonly determines the ratio of the fluxions of quantities', says MacLaurin, 'but we shall treat more fully of his method afterwards; and, since there have been various objections made against this doctrine, we shall demonstrate its principal propositions immediately from the axioms' (pp. 101–102).

In Chapter II, 'Of the fluxions of plane rectilinear figures', MacLaurin proves three propositions, as usual with a classical double *reductio ad absurdum* and no appeal to infinite processes or to infinitely small quantities, leading up to finding the fluxion of a rectangle, i.e., of a product. He finds the fluxion of a parallelogram of constant height and flowing base, reduces finding the fluxion of a triangle to finding the fluxion of an auxiliary parallelogram, and then finds the fluxion of a rectangle both sides of which flow by dividing it into two triangular areas whose fluxions are found by the previous proposition. He develops corollaries about the area of the triangle that grows with a uniform acceleration as the base flows uniformly, and relates this to the uniformly accelerating effect of a constantly applied force like gravity. All this is done without infinitesimals, infinite processes or actual infinites, in the manner of the ancients, and MacLaurin says that these theorems constitute the foundation of the direct method of fluxions.

As Newton observed in the *Principia*, from the fluxion of a product one can easily derive the fluxion of an arbitrary integral power (§5.4); but MacLaurin derives that fluxion in another way as well. In his Chapter V, 'Of the fluxions of Quantities that are in a continued geometrical Progression, the first term of which is invariable', MacLaurin states the following proposition (p. 156):

> Proposition VII. The fluxion of any term *AN* of a geometrical progression, the first term of which is invariable, is to the fluxion of the second term *AP* in a ratio compounded of the ratio of those terms and the ratio of the number of terms which precede *AN* to unit.

That is, if

$$AS : AP :: AP : AL :: AL : AM :: AM : AN :: \text{etc.} \tag{1}$$

Then

$$\frac{\textit{fluxion of } AN}{\textit{fluxion of } AP} = \left(\frac{AN}{AP}\right)\left(\frac{4}{1}\right). \tag{2}$$

This is demonstrated by means of the construction shown in Figures 1 and 2. If *AS* is perpendicular to *AP*, and the angles at P, L, M etc. are right angles, then by similar triangles, *AS* : *AP* :: *AP* : *AL* :: *AL* : *AM* :: *AM* : *AN* :: etc. MacLaurin keeps S fixed and lets P move along the axis while all the angles remain right, and uses Euclidean geometry to prove propositions about the relations among the increments of *AP*, *AL*, *AM*, etc., traced out by P, L, M, etc.

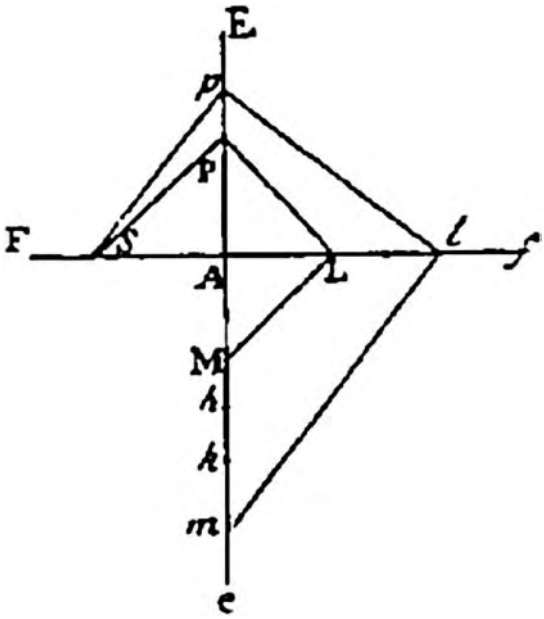


Figure 1. MacLaurin's Figure 40.

That the fluxion derived in the proposition is in fact equivalent to the fluxion of an integral power can be seen by letting $AS = a$, and $AP = ax$ (so that the index of the progression is x). Then

$$AS : AP :: AP : AL :: AL : AM :: AM : AN :: \text{etc.} \tag{3}$$

becomes

$$a : ax :: ax : ax^2 :: ax^2 : ax^3 :: ax^3 : ax^4 :: \text{etc.} \tag{4}$$

and

$$\frac{\textit{fluxion of } AN}{\textit{fluxion of } AP} = \left(\frac{AN}{AP}\right)\left(\frac{4}{1}\right) \tag{5}$$

becomes

$$\frac{\textit{fluxion of } ax^4}{\textit{fluxion of } ax} = \left(\frac{ax^4}{ax}\right)\left(\frac{4}{1}\right) = 4x^3. \tag{6}$$

This proposition is also put to extensive use in Chapter VI, 'Of logarithms, and the fluxions of logarithmic quantities'. MacLaurin defines logarithms after the manner of John Napier, and develops the fluxions of and relations among logarithmic quantities. He notes the relation to exponentials and to the hyperbolic area, and discusses logarithms of different moduli. He says that areas of all conics are now reduced to the measures of lines, angles (i.e., trigonometric functions), or ratios (i.e., logarithmic functions), and that all fluents in

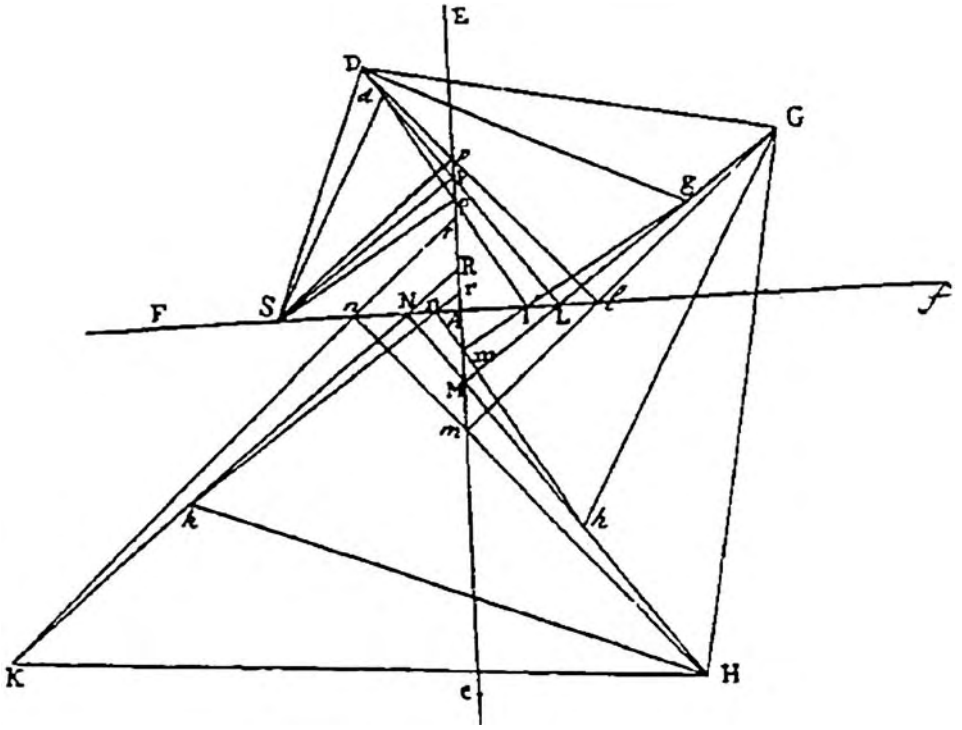


Figure 2. MacLaurin's Figure 39.

the method of fluxions should be reduced to these measures as much as possible (rather than, as was usual, having immediate recourse to infinite series).

3 EXTREMA AND INFLECTIONS

Chapter IX treats 'Of the greatest and least ordinates, of points of contrary flexion and reflexion of various kinds, and of other affections of curves that are defined by a common or by a fluxional equation'. Here and in Chapter V of Book II, MacLaurin gives the most complete treatment of extrema and points of inflection to be found in any 18th-century British work, and makes the first enunciation of the complete fluxion test for such points, using all orders of fluxions, and not simply identifying critical points with a first or second fluxion test, which must then be investigated further by testing neighboring points (pp. 693–695). Graphing the fluent and the fluxion over the same base, he observes the relationship between tangents and areas, and he observes that not only the value of the fluxion as its curve crosses a particular ordinate, but also the way in which the curve of the fluxion crosses the ordinate gives information about the curve of the fluent at that ordinate. Information about this fluxional curve is in turn revealed by curves of higher fluxions, and so forth.

MacLaurin also classifies and gives rules for distinguishing various types of inflections, cusps, and maxima and minima. He continues his treatment of extrema and inflections in

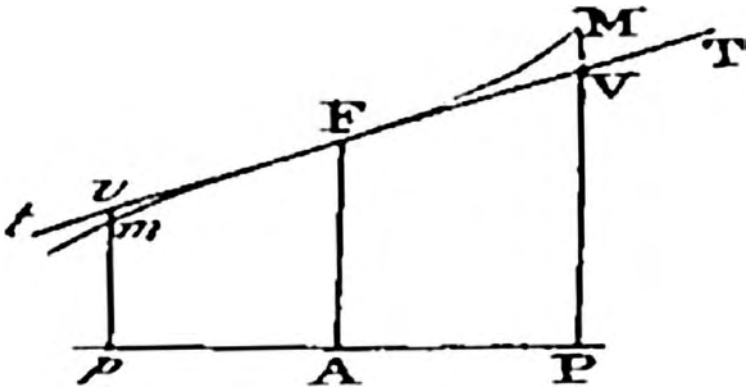


Figure 3. MacLaurin's Figure 319.

Book II, using the MacLaurin Series. In this more algebraic treatment, MacLaurin says that if the ordinate $AF = E$, $AP = x$ (see Figure 3), and the base is supposed to flow uniformly, then the ordinate

$$PM = E + \dot{E}/\dot{x} + \ddot{E}x^2/2\dot{x}^2 + \dddot{E}x^3/6\dot{x}^3 + \&c. \tag{7}$$

The equivalence of this form to the modern version of the MacLaurin series will be apparent by placing the origin at A and letting AF be the y-axis, so that E is $f(0)$, and understanding $\dot{E}/\dot{x}$, $\ddot{E}/\dot{x}^2$, $\dddot{E}/\dot{x}^3$, etc. to be equal to $f'(0)$, $f''(0)$, $f'''(0)$, etc. Likewise the ordinate

$$pm = E - \dot{E}x/\dot{x} + \ddot{E}x^2/2\dot{x}^2 - \dddot{E}x^3/6\dot{x}^3 + \&c. \tag{8}$$

(the signs alternate because x is understood to be the absolute distance AP, i.e., it is always positive).

Now if the first fluxion $\dot{E} = 0$, then

$$PM = E + 0 + \ddot{E}x^2/2\dot{x}^2 + \dddot{E}x^3/6\dot{x}^3 + \&c., \tag{9}$$

and

$$pm = E - 0 + \ddot{E}x^2/2\dot{x}^2 - \dddot{E}x^3/6\dot{x}^3 + \&c. \tag{10}$$

Therefore, when AP and Ap are small, the ordinates PM and pm will both exceed AF when $\ddot{E}$ is positive, and will both be less than AF when $\ddot{E}$ is negative. If $\ddot{E}$ is also zero, but $\dddot{E}$ is not, then a consideration of the series shows that one of the ordinates PM and pm will be greater and the other less than AF. This line of argument leads to the following rule (pp. 694–695):

> In general, if the first fluxion of the ordinate, with its fluxions of several subsequent orders, vanish, the ordinate is a maximum or minimum, when the number of all those fluxions that vanish is 1, 3, 5, or any odd number. The ordinate is

> a minimum, when the fluxion next to those that vanish is positive; but a maximum when this fluxion is negative. [. . .] But if the number of all the fluxions of the ordinate of the first and subsequent successive orders that vanish be an even number, the ordinate is then neither a maximum nor minimum.

Similar rules are developed for points of inflection.

4 LIMITS OF SERIES, AND THE EULER–MACLAURIN THEOREM

MacLaurin treats 'Of the Asymptotes of curve Lines, the Areas Bounded by them and the Curves, the Solids Generated by those Areas, of Spiral Lines, and of the limits of Sums of Progressions' in Chapter X. He defines asymptotes and the continual approach to a limit, and he discusses how to determine whether the area between a curve and its asymptote is limited or not. He extends this to the volumes of solids of revolution. He notes the analogy between areas under asymptotic curves and the sums of infinite sequences. He gives rules for sums of differences and higher order differences, and he shows how to derive numerous summable sequences from other sequences. His definition of the limit of the sum of an infinite progression is as follows (p. 289):

> As a right line, or figure, may increase continually and never amount to a given line, or area; so there are progressions of fractions which may be continued at pleasure, and yet the sum of the terms be always less than a certain finite number. If the difference betwixt their sum and this number decrease in such a manner, that by continuing the progression it may become less than any fraction how small soever that can be assigned, this number is the limit of the sum of the progression, and is what is understood by the value of the progression when it is supposed to be continued infinitely. These limits are analogous to the limits of figures which we have been considering, and they serve to illustrate each other mutually.

An admirable way in which they 'illustrate each other mutually' is in what has come to be called the 'Euler–MacLaurin integral test' for the convergence of a series [Mills, 1985].

Chapter XI is 'Of the curvature of Lines, its Variation, and the different kinds of contact of the Curve and the Circle of curvature, The Caustics by Reflexion and Refraction, the Centripetal forces, and other problems that have a Dependence upon the Curvature of Lines'. MacLaurin considers the properties of the curvature of numerous curves, and applies curvature to the topics of the title of this chapter, especially to centripetal forces, including planetary motion and the three body problem, and motion in a void and in a resisting medium. The concluding paragraph of this chapter asserts that the principal propositions of the method of fluxions have now been deduced from plain axioms, and the method is now as certain and evident as the common geometry. 'We have insisted on it at so great length', says MacLaurin (pp. 411–412),

> chiefly because a full account of the manner in which the principal propositions of the method of fluxions are demonstrated by it, may be of use for removing several objections that have been lately urged against this doctrine; which has

> been represented, as depending on nice and intricate notions; while it has been insinuated, that they who have treated of it have been earnest rather to go on fast and far, than solicitous to set out warily, and see their way distinctly. But we now proceed to the more concise methods by which the fluxions of quantities are usually determined.

5 THE METHOD OF INFINITESIMALS, AND NEWTON'S PRIME AND ULTIMATE RATIOS

In Chapter XII, 'Of the Method of Infinitesimals, of the Limits of Ratios, and of the General Theorems which are Derived from this Doctrine for the Resolution of Geometrical and Philosophical Problems', MacLaurin explains the success of the method of infinitesimals in spite of its apparently illegitimate neglecting of terms, and explains away the apparent use of such methods by Newton. In the method of fluxions, he says, if the generating motion is uniform, the fluxion is measured by the increment acquired in a given time. If it is accelerated, the increment is resolved into two parts; that which alone would have been generated if the motion had not been accelerated, and that which was generated in consequence of the acceleration. Terms neglected in the method of infinitesimals are of the latter sort, and this is why no error results.

MacLaurin states that however safe this method may be, it is not appropriate 'to admit infinitely little quantities, and infinite orders of infinitesimals, into a science that boasts of the most evident and accurate principles as well as of the most rigid demonstrations'. To avoid this, he has founded the Method of Fluxions 'on more unexceptionable postulata' in the preceding chapters. Newton avoided infinitesimals, MacLaurin claims, by considering simultaneous finite increments, and investigating the limiting ratios to which proportions of these increments approach, as they are supposed to decrease together until they vanish. Newton determines this limit by reducing the expression of the ratio of the finite increments to simplest terms, so that part of the expression is seen to be independent of the increment; 'then by supposing the increments to decrease till they vanish, the limit readily appears' (pp. 420–421). MacLaurin says that Newton has been accused of shifting the hypotheses in such demonstrations, but that this is unjust. He assumes an increment, forms a ratio, and investigates 'the ratio of those increments at any term of the time while they had a real existence, how this ratio varied, and to what limit it approached, while the increments were continually diminished'. Letting them vanish 'is a concise and just method of discovering the limit which is required'. The prime or ultimate ratio 'strictly speaking', is not the ratio of any real increments whatever, but the limit of the variable ratio of the increments (p. 422).

Book I concludes with Chapter XIV, 'Of the ellipse considered as the section of a cylinder. Of the Gravitation towards Bodies, which results from the Gravitation towards their Particles. Of the figure of the Earth, and the Variation of Gravity Towards it. Of the ebbing and Flowing of the Sea, and other inquiries of this nature'. The initial motivation came from a prize problem on tides set for 1740 by the Paris *Académie des Sciences.* MacLaurin was one of the winners, with an essay which contained theorems on 'level surfaces' (now called 'equipotentials') of attraction of the Earth to an external point lying upon an

axis. He expanded upon the essay in this chapter; his results contributed notably to the development of planetary mechanics (compare §18.5 on Laplace). Besides MacLaurin's pioneering work on the attraction of ellipsoids, this chapter is interesting for the treatment of conics as projections of a circle, with their invariant properties established in the case of a circle and transferred by projection to the general conic.

6 BOOK II

Having treated fluxions geometrically in the first 574 pages of his treatise, so as to facilitate his foundation of the method 'in the manner of the ancients', MacLaurin devotes the last 180 pages to an algebraic treatment in Book II, 'On the Computations of the Method'. He does not, however, leave foundational questions behind him, or justify his algebraic procedures solely on their geometric interpretation, but shows that an algebraic interpretation of the method of fluxions can also be founded on the same sorts of demonstrations, without infinitesimals or appeals to infinite processes.

In Chapter I MacLaurin addresses Berkeley's assertion that mathematicians hide unclear concepts behind distinct notation. After discussing the generalizing power of algebra versus geometric constructions, MacLaurin says (p. 576):

> It may have been employed to cover, under a complication of symbols, abstruse doctrines, that could not bear the light so well in a plain geometrical form; but, without doubt, obscurity may be avoided in this art as well as in geometry, by defining clearly the import and use of the symbols, and proceeding with care afterwards.

To this end, MacLaurin says that in the algebraic treatment of the method, quantities are no longer considered as generated by motion, but the respective rates with which they increase or decrease when they are supposed to vary together are ascertained. 'By the fluxions of quantities we shall therefore now understand, any measures of their respective rates of increase or decrease, while they vary (or flow) together'. To deal with such fluxions defined in this way, MacLaurin develops algebraic analogs of the four geometric axioms he presented at the very beginning of the treatise. He then uses these principles to prove six propositions giving the fluxions of the square, of the nth power, the (n/m)th power, of products and quotients, and of logarithms, as in his geometric demonstrations, using no infinitesimal quantities nor appeals to infinite processes. He proceeds by establishing a finite inequality that holds between expressions involving intervals of the variables and the proposed value of the fluxion, and then showing that supposing the true value of the fluxion to differ from that proposed will lead to a contradiction with this inequality when the interval is taken sufficiently small, amounting in essence to what we would call an epsilon–delta demonstration (p. 584).

In Chapter III, 'Of the analogy between circular arches and logarithms, or reducing fluents to these, or to hyperbolic and elliptic arches, or to other fluents of a more simple form; when they are not assignable in finite algebraic terms', MacLaurin presents a collection of integration techniques (of course, he did not use that term). He reduces fluents to areas of conics, and to arc lengths of conics. He discusses change of variable, and makes extensive

use of partial fractions. And he treats fluents not expressible even by hyperbolic or elliptic lengths, which can be expressed by sums or differences of such quantities. More 'integration techniques' are presented in Chapter IV, 'Of the area when the ordinate and base are express'd by fluents; of computing the fluents from the sums of progressions, or the sums of progressions from the fluents, and other branches of this method'. Included are what we call the 'chain rule' and 'integration by parts'; there are numerous theorems and examples dealing with sequences and interpolation.

In the final Chapter V, 'Of the general rules for the resolution of problems', MacLaurin applies the techniques of Book II to problems and theorems demonstrated geometrically in Book I. These include extrema, L'Hôpital's rule, points of contrary flexure, cuspids, centripetal forces and trajectories, motion on a cycloid, the catenary, solids of revolution, spherioids and the attraction and shape of the Earth, centers of gravity and of oscillation, hydrodynamics, isoperimetric problems, optimum angle of vanes and sails, optimum course of ships, reducing the order of a fluxional equation, the elastica, vibrating chord, and solid of least resistance.

7 SUMMARY REMARKS

Although MacLaurin's treatise receives perfunctory praise in all the standard histories of mathematics, it is often represented as extremely prolix and difficult to understand. His defense of Archimedes against similar charges is as applicable to his own book (pp. 35–36):

> His method has been often represented as very perplexed, and sometimes as hardly intelligible. But this is not a just character of his writings, and the ancients had a different opinion of them. He finds it necessary indeed to premise several propositions to the demonstration of the principal theorems; and on this account his method has been excepted against as tedious. But the number of steps is not the greatest fault a demonstration may have; nor is this number to be always computed from those that may be proposed in it, but from those that are necessary to make it full and conclusive. Besides, these preliminary propositions are generally valuable on their own account, and render our view of the whole subject more clear and compleat.

To further clarify our view of the whole subject, and to demonstrate that these foundations will support the entire received structure of the method of fluxions and the calculus, MacLaurin applied the method as he had developed it across the entire range of 18th-century mathematics. Besides the basic quadratures, cubatures, maximum/minimum, tangents, and rates of change problems of the elementary method, he treated all the challenge problems and curves of such interest to the continental analysts. He treated mechanics, elastic and inelastic collision, logarithms and other sequences, curvature and its variation, caustics, cycloids, the rainbow, centripetal forces, trajectories under all sorts of forces, motion on various curves, celestial mechanics, the shape of the earth and motion of the tides, centers of gravity and of oscillation, pendular motion, hydrodynamics, the catenary, tautochrone, and brachistochrone, the general isoperimetric problem, and projective properties

of the conics. In Book II he presented 'the computations of the method', developing and demonstrating techniques for finding fluxions and fluents of expressions with much greater facility than one might have expected of the Newtonian notation, and applying these techniques to the problems of the first book. In the process he treated most of the 'integration techniques' encountered in a modern undergraduate calculus course, fluents expressible in closed form only as elliptic and hyperbolic curve lengths or their sums and differences, and infinite series with consideration of convergence.

8 IMPACT

The *Treatise* was generally cited by British fluxionists as the definitive answer to Berkeley's criticism, but MacLaurin had accomplished much more than this. Judith Grabiner has described MacLaurin's influence on the Continental analysts in detail. MacLaurin's work was cited with admiration by Lagrange, Euler, Clairaut, d'Alembert, Laplace, Legendre, Lacroix, and Gauss. The influence of MacLaurin's use of the algebra of inequalities as a basis for his limit arguments can be seen in d'Alembert, L'Huilier, Lacroix and Cauchy. MacLaurin corresponded at length with Clairaut about the attraction of ellipsoids, and the latter in his *La figure de la terre* (1743) acknowledges his debt; MacLaurin's influence on this subject can be seen also in d'Alembert, Laplace, Lagrange, Legendre, and Gauss. Of especial note are his contributions in Chapter XIV to the theory of 'level surfaces' (his name: we now call them 'equipotential'), which partly draw upon earlier work. He showed that if we assume the inverse square law of attraction, then the ratio of the attractions of two homogenous confocal ellipsoids to an external point lying along a principal axis was the same as the ratio of their masses. His use of geometrical reasoning contrasts strikingly with contemporary studies of the attraction of ellipsoids that were carried out by Clairaut, who used techniques from mathematical analysis [Greenberg, 1995, 412–425, 587–601].

In addition, MacLaurin's use of infinite series in the analysis of functions, especially with the Euler–MacLaurin formula, was known to Euler, Lagrange, and Jacobi; while his reduction of fluents to elliptic or hyperbolic curve length was used by d'Alembert and extended by Euler, and Euler influenced Legendre's work on elliptic integrals [Grabiner, 1997, 400–403].

BIBLIOGRAPHY

Berkeley, G. 1734. *The analyst*, London: J. Tonson in the Strand. [See §8.]

Cajori, F. 1919. *A history of the conceptions of limits and fluxions in Great Britain, from Newton to Woodhouse*, Chicago and London: Open Court.

Grabiner, J. 1997. 'Was Newton's calculus a dead end? The Continental influence of Maclaurin's treatise of fluxions', *American mathematical monthly*, *104*, 393–410.

Grabiner, J. 2002. 'MacLaurin and Newton: The Newtonian style and the authority of mathematics', in C.W.J. Withers and P. Wood (eds.), *Science and medicine in the Scottish Enlightenment*, Edinburgh: Tuckwell Press, 143–171.

Greenberg, J.L. 1995. *The problem of the Earth's shape from Newton to Clairaut*, Cambridge: Cambridge University Press.

Guicciardini, N. 1989. *The development of the Newtonian calculus in Britain 1700–1800*, Cambridge: Cambridge University Press.

Mills, S. 1985. 'The independent derivations by Leonard Euler and Colin MacLaurin of the Euler–MacLaurin summation formula', *Archive for history of exact sciences, 33*, 1–13.

Turnbull, H.W. 1947. 'Colin Maclaurin', *American mathematical monthly*, *54*, 318–322.

Turnbull, H.W. 1951. *Bicentenary of the death of Colin MacLaurin*, Aberdeen: Aberdeen University Press.

Tweedie, C. 1915, 1919, 1921. 'A study of the life and writings of Colin MacLaurin', *The mathematical gazette*, *8*, 132–151; *9*, 303–306; *10*, 209.

CHAPTER 11

JEAN LE ROND D'ALEMBERT, *TRAITÉ DE DYNAMIQUE* (1743, 1758)

Pierre Crépel

This book, D'Alembert's magnum opus, was one of the first to give a unified view of mechanics. It started out from a minimum of principles, one of which came to be named after him.

First publication. Paris: David, 1743. xxvi + 186 pages + plates (+ 8 unnumbered pages).

Second edition. Paris: David, 1758. xl + 272 pages + plates (+ 10 unnumbered pages) + lunar table.

Posthumous reprint. Paris: Fuchs, 1796. [Based upon the 1758 edition, with some minor modifications. Contains an addition to art. 176, already published in his *Opuscules mathématiques*, vol. 1 (1761).]

Photoreprints. 1st ed. Brussels: Culture et Civilisation, 1967. 2nd ed. Paris: Gabay, 1990 [with some modifications].

Critical edition in preparation, as his *Œuvres complètes*, series 1, volume 2, Paris: CNRS-Editions.

German translation of the 1st edition. *Abhandlung der Dynamik* (trans. and ed. A. Korn), Leipzig: Engelmann, 1899 (*Ostwalds Klassiker der exakten Wissenschaften*, no. 106). [Photorepr. Thun & Frankfurt: Harri Deutsch, 1997.]

Russian translation. *Mekhanika* (trans. and ed. Ergochin), Moscow: Nauka, 1950.

Manuscripts. Apparently lost, but some relevant ones printed in *Opuscules*, vol. 9 (1783).

Related articles: Newton (§5), Lagrange on mechanics (§16), Montucla (§21), Laplace on astronomy (§18), Thomson and Tait (§40).

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

1 BIOGRAPHY OF D'ALEMBERT

A natural son of the chevalier Destouches and Mme. De Tencin, D'Alembert was born on 16 or 17 November 1717 and was placed (rather than abandoned) on the steps of the church of Saint-Jean-le-Rond in Paris—whence his given name, although much later he preferred 'Daremberg', then 'Dalembert' or 'D'Alembert'. He followed his secondary studies at the *Quatre-Nations* College in Paris, and later studied law and probably a little medicine. His first memoir was submitted to the *Académie des Sciences* in Paris in 1739, and he became a member of that institution in 1741.

The *Traité de dynamique* was D'Alembert's first major work, to be followed by many others in the 1740s and early 1750s. He was co-editor with Dénis Diderot of the *Encyclopédie*, for which he wrote the introduction (1751) and around 1700 articles, mainly scientific, the majority of them before the work was banned following his article 'Genève' in 1758–1759. He was appointed to membership of the *Académie Française* in 1754 and quickly became second to Voltaire in the group for *philosophes*. In 1772 he became permanent secretary of this academy (but not that for sciences). He died of gall-stones on 29 October 1783. On his life, see Hankins [1970].

2 SCIENTIFIC WORKS

A multifaceted enterprise, the edition of the complete works currently being prepared will consist of around forty volumes of about 700 pages each. It is divided into five series of comparable size, as follows. Series I: mathematical treaties and memoirs (1736–1756); Series II: contributions to the *Encyclopédie*; Series III: mathematical notes and memoirs (1757–1783); Series IV: mixed, history, literature, philosophy; Series V: correspondence. The period covered by the first series (roughly as far as the *Encyclopédie*) can be regarded as the golden age for mathematics. In about a decade, the author published six celebrated treatises, beginning with the *Traité de dynamique* (1743), as well as two others on fluids, one on the cause of winds, one on the precession of the equinoxes; and, finally, one on the 'system of the world' in three volumes covering various branches of astronomy and the shape of the Earth. He published a similar number of important memoirs, notably on vibrating strings and what we call the fundamental theorem of algebra.

Not to be omitted from his scientific output are D'Alembert's articles for the *Encyclopédie*, which not only played a popularizing role but sometimes contained new research or clarified that of others.

Finally, in the period extending from the mid 1750s to his death, despite having fallen out with the academies of Berlin and Paris, he nevertheless drafted at least 5000 pages of mathematics, mainly in the form of 'Notes'. In these he returned to a number of earlier subjects, but often adding remarkable ideas that posterity has underestimated up to now. We note in conclusion that scientific ideas are scattered about in various philosophical writings and in his correspondence, especially in the 170 or so letters exchanged with J.L. Lagrange (1736–1813) between 1759 and 1783.

We thus have a legacy that is abundant and also diversified: differential and integral calculus, mechanics, hydrodynamics, astronomy, optics, the shape of the Earth and probability

theory. The name of D'Alembert survives in the mathematics of today: the 'Dalembertian' for the wave equation, D'Alembert's principle in mechanics, D'Alembert's paradox in hydrodynamics, and so on. We mention that, at least in France, the fundamental theorem of algebra is generally referred to as 'the D'Alembert–Gauss theorem'. Likewise, the rule for the convergence of numerical series, whereby the ratio u_{n+1}/u_n of two consecutive terms is bounded by a number less than unity, is known to students as 'D'Alembert's ratio test'.

However, D'Alembert's writings have some puzzling characteristics. He often follows his thoughts and constantly criticizes his contemporaries; his writings are poorly structured and not pedagogical, much less polished than those of Leonard Euler (1707–1783) for example; he has his own way of getting to the bottom of things whose profundity can usually be grasped only at a second reading. He has often been judged severely by history, although this tendency has been reversed over the last two decades.

3 CONTENTS OF THE *TRAITÉ DE DYNAMIQUE*

We first describe the first edition, then the differences introduced in the second. Their contents are summarized in Table 1.

The book was published in 1743 by David, the great bookselling and printing house, in the classical binding approved and favoured by the King. It was presented to the *Académie des Sciences* on 22 June and given a favourable review by the commissioners P.L. Maupertuis and F. Nicole. It includes a letter to Count de Maurepas, a 26-page preface summarizing and commenting on the main general ideas, and finally the body of the text with all its trimmings (table of contents, plates, corrections, extract from the records of the *Académie des Sciences*, royal favour).

Contrary the author's current custom, the book is clearly structured. Following the definitions and preliminary notions (pp. 1–2), the first Part is entitled 'General laws of motion and equilibrium of bodies' (pp. 3–48). It consists of three chapters, each of which is subdivided into articles numbered continuously. These chapters represent the three great principles on which dynamics is based: I. 'On the force of inertia' (arts. 2–20); II. 'On composite motion' (arts. 21–26); III. 'On motion destroyed or changed by obstacles' (arts. 27–49). This last chapter contains, in particular, the theory of equilibrium.

The second Part, which is much larger (pp. 49–186), is entitled 'A general principle for finding the motion of many bodies that act on each other in an arbitrary way, with many applications of this principle'. It consists of four chapters of disparate length and status: I. 'Exposition of the principle' (art. 50); II. 'Properties of the centre of gravity of many bodies combined, deduced from the preceding principle' (arts. 51–72); III. 'Problems illustrating the application of the preceding principle' (arts. 73–153); IV. 'On the principle of conservation of live forces' (arts. 154–175). What is today called 'D'Alembert's principle' constitutes the single article of Chapter I. The rest of the second Part consists of what the author calls 'applications'.

While the book is structured in a straightforward way in terms of definitions, lemmas, theorems, laws, corollaries, remarks and problems, it is nevertheless rather difficult to read,

Table 1. Comparison of the editions of 1743 and 1758 of D'Alembert's book.

Part of the edition	**1743: articles/pages**	**1758: articles/pages**
Title page	'D'Alembert membre de l'Académie des sciences. Chez David, l'aîné'	'D'Alembert membre de nombreuses académies. Chez David'
Dedication	To the Count Maurepas	To the Count d'Argenson
Permissions, privileges	*Académie des Sciences*, 22 June 1743	*Académie des Sciences*, 26 April 1758
Warning	None	Differences between the editions
Preliminary material	'Preface', p. j	Augmented 'Preliminary discourse', pp. j–xxxv
Part 1. *'General laws'*.		
Ch. I. 'Inertia'.	2–20/3–22	2–27/3–34
Ch. II. 'Compound motion'.	21–26/22–31	28–33/35–44
Ch. III. 'Destroyed motion'.	27–49/31–48	34–59/44–71
Part 2. *Bodies which act upon each other*.		
Ch. I. Principle' of D'Alembert.	50/49–52	60–61/72–75
Ch. II. 'Centre of gravity'.	51–72/52–69	62–86/75–96
Ch. III. 'Applications'.		
Sec. I. 'Bodies that push by threads or by rods'.	73–114/69–122	87–144/96–186
Sec. II. 'Bodies that oscillate'.	115–120/122–129	145–150/186–200
Sec. III. 'Bodies that move freely on rods'.	121–124/129–138	151–155/200–211
Sec. IV. 'Bodies that push or collide'.	125–153/138–169	156–185/211–252
Ch. IV. 'Conservation of live forces'.	154–175/169–186	186–207/252–272
Further material.		A lunar table.
Plates; errata.	I–IV; yes	I–V; no

for both contemporary and modern readers. This is due to the style of the author, the form of the figures, the notation (where the same letter often denotes different things), the relationship between the text and the diagrams, the use of differential notation, and a personal conception of vocabulary (words such as 'force' and 'power' do not have the same meanings as today, and even seem to designate physical concepts that we regard as different). Finally, it must be said that D'Alembert makes little attempt at pedagogy.

'D'Alembert's principle' (Chapter I of Part II) plays a pivotal role in the book: Part I paves the way for it and Chapters II–IV of Part II consist of applications [Fraser, 1985]. Moreover, it is this principle that posterity has universally accepted as one of D'Alembert's main contributions to science. To explain it, we begin by quoting the first sentence of the chapter:

> Bodies act as one another in only three different ways that are known to us: by immediate impulse, as in the case of an ordinary impact; by the interposition between them of some body to which they are attached; by virtue of mutual attraction, as in the Newtonian system of the Sun and the Planets.

Estimating that the effects of the last type of action have been sufficiently well examined, the author restricts his attention to the first two. The 'General problem' and the solution that follow in Chapter I constitute the famous principle. This technical passage is not so easy to read, but the basic idea is explained, admittedly without too much emphasis, in the preface, as follows:

> Just as the motion of a body which changes direction can be regarded as composed of the motion it had originally and a new motion that it has acquired, so the motion that the body had originally can be regarded as composed of a new motion that it has acquired, and another that it has lost. It follows from this that the laws of a motion changed by obstacles depend only on the laws of the motion destroyed by these obstacles. For it is obviously sufficient to decompose the motion of the body before meeting an obstacle into two other motions, one of which is unaffected by the obstacle while the other is annihilated.

In modern notations, let the vector n denote the new motion, a the old motion, r the acquired motion, and p or $-r$ the motion lost. Then the above quotation reduces to the statement that $n = a + r$, or $a = n + p$, where the obstacle does not affect n and annihilates p.

D'Alembert deduces from it that the determination of all motions reduces to applying the principle of equilibrium and that of composite motion. That is why it is often said that D'Alembert's principle reduces dynamics to statics. The simplest example is that of a body without elasticity obliquely striking a fixed impenetrable wall: the only component of motion preserved after the impact is that parallel to the wall, the component perpendicular to the wall being destroyed (Part I, Chapter III). A typical theorem from Chapter II is as follows: 'The state of motion or rest of the centre of gravity of many bodies does not change under the mutual action of these bodies provided that the system is entirely free, that is, it is not subject to motion around a fixed point'.

Chapter III, which is by far the longest and takes up more than half of the book, contains a detailed treatment of 14 problems, divided into Sections as follows: I. 'Bodies pulled by wires or rods' (Problems I–VI); II. 'Bodies moving in the plane' (Problem VII); III. 'Bodies acting on one another via wires along which they can run freely' (Problem VIII); IV. 'Bodies which move or collide' (Problems IX–XIV). These problems are treated at unequal length, some being more famous than others. For example, Problem V (arts. 98–112 in the first edition), on the period of oscillation of a composite pendulum, formed a part of the immediate prehistory of the problem of the vibrating string.

In Chapter IV, the author emphasizes the fact that, contrary to the Bernoullis, he does not assume the conservation of live forces, but that it can be deduced from his principle and methods. He states that he gives, 'if not a general proof for all cases, at least a sufficient number of principles for finding the proof in each particular case'. He sketches these proofs for bodies on wires or rods in the case of elastic impacts, and for fluids.

4 ON THE SECOND EDITION

Let us turn to the differences between the first and second editions. They are easily identified as D'Alembert lists them in the foreword of the 1758 edition.

Beginning with the context, this time the letter is addressed to the Count d'Argenson, not to Maurepas. The work was again presented to the *Académie des Sciences*; the report by Etienne Bézout (1730–1783) and E. Montigny is dated 26 April 1758.

The preface was renamed 'Preliminary discourse'. It is little changed, but the following words are added: 'some reflections on the question of live forces and an examination of another important question posed by the Royal Academy of Sciences of Prussia as to whether the laws of Statics and Mechanics are indeed necessary or contingent?'.

D'Alembert lists on one page the various places where the text has been improved, simplified or enriched, as in the case of the impact of resilient bodies, for example. He then draws attention to the 61 notes made by Bézout at his request 'to make the work accessible to a greater number of readers than the first edition', which is a polite way of saying that most people had let him know that they had not understood very much!

Finally, it is interesting to note that in 1758, D'Alembert was already planning the further development of points of difficulty in the treatise. These writings, already in draft form, would constitute the first five memoirs in volume I of his *Opuscules mathématiques,* published in 1761, as follows: the motion of a body turning around a moving axis (Memoir 2); additions to the essay on the resistance of fluids (Memoir 4); a theory of oscillation of floating bodies (Memoir 3); replies to Daniel Bernoulli (1700–1782) and Euler on vibrating strings (Memoir 1); and another proof of the principle of composition of forces (Memoir 5). The nine volumes of *Opuscules* (1761–1783) contain many other memoirs connected with the treatise, as we shall see below.

An unexpected curiosity is the addition of a lunar table, apparently without any relationship to the subject. This has to do with a custom that was prevalent in the 18th century, due to delays in publication, the price of books and publishing difficulties: when the author of a treatise had drafted a research memoir on another subject without having arranged for its publication, he would often insert it as an appendix in the earlier treatise in the course of printing. This table was suppressed in the 1796 edition, and again in the 1990 reprinting.

As D'Alembert points out in the first line of the foreword, 'this second edition is augmented by more than a third'. One must, however, avoid an error of misinterpretation: this refers essentially to the same work and not a complete rewrite.

5 THE PLACE OF THE *TREATISE* IN THE WORK OF THE AUTHOR

This question is more difficult than one might think. D'Alembert's work had numerous facets, and their unity can be interpolated in different ways. It has recently been shown, for

example, that the relationship between pure mathematics and the physical sciences in his work was a very subtle one [Firode, 2001].

The *Traité de dynamique* nevertheless permeates a large part of his work, quite explicitly so in the case of the *Traité des fluides*, which was published in the following year (1744) as a continuation of the earlier work. It is also true that the majority of his work on the physical sciences make use of D'Alembert's principle, as do the *Recherches sur la cause des vents* of 1747 of those on the *Précession des equinoxes* two years later, or the *Essai sur la résistance des fluides* of 1752. It may be said that, even if these treatises are bursting with other interesting discoveries, they are also 'applications' of the treatise.

At another level, the preface and the basic concepts, which form the essence of the book and are bereft of the more tedious formulae, pervade the author's 'popular' writings. Some quite long passages can be found 'pasted into' his articles in the *Encyclopédie* and in the *Mélanges d'histoire, de literature et de philosophie* (1753 and later).

There is another aspect, largely ignored by historians of science, and that is that D'Alembert spent part of the 40 years between the publication of the treatise and his death in commentating on, improving and somewhat extending every aspect of it, in (parts of) the memoirs in the *Opuscules mathématiques* (which include an unpublished ninth volume whose publication was prevented by his death). Thus memoirs 2, 21 and 22 of volumes I (1761) and IV (1768) extend the remarkable theory of the motion of arbitrary rotating bodies (possibly without a fixed point) already developed in the researches on the precession of equinoxes. Also seven or eight circulated memoirs contain explanations, modifications, variants and alternative proofs of the three great fundamental principles: inertia, composition of motions and equilibrium. There is also further discussion of the 'problems' in the second part of Chapter III, such as those involving elasticity, impacts and impulses. Finally, in the numerous memoirs on fluids, in particular 4 (volume I) and above all art. IV of 51 (volume VI, 1773) and 57 (volume VIII, 1780), one can find new proofs prompted by questions of J.-C. Borda on the conservation of live forces, a problem already addressed at the end of the first edition of the treatise (Chapter IV of Part 2).

To appreciate D'Alembert's thinking on the foundations and methods of mechanics, it is therefore necessary to take into account all of his work, not just the treatise. But there is no need to be carried away by an excess of erudition: in the main, these later memoirs confirm the thinking of the author, reinforcing it and modifying minor aspects, but the great ideas are already in the first edition of the treatise (1743).

6 POSTERITY OF THE *TREATISE*

When examining the impact of the treatise on both contemporaries and the centuries that followed, one needs to exercise a certain caution. The book did not by any means pass unnoticed: its appearance was an event, a fact to which the registers of the *Académie des Sciences*, the journals and correspondence bear abundant witness. But who actually read and understood it at the time?

We must not impose our preoccupation with the history of science on the scholars of the 18th century. Daniel Bernoulli, Euler, Alexis Clairaut (1713–1765) and Lagrange were much too creative geniuses to have the patience to plough through a book like this treatise,

at least two thirds of which consists of rather tedious applications that are not structured in a pedagogical way. It is clear that they read the preface, the essentials of the first part and the beginning of the second, where the famous principle appears. As to the rest, we shall never know: probably rather little, and in any case not continuously. Such scholars as those always read a book with their own projects in mind and rarely put themselves in the place of their colleague: they content themselves with snapping up those ideas that might prove to be useful in their own work. Nevertheless, they all grasped the importance, the essential message, of the book: in particular, the *Méchanique analitique* (1788) of Lagrange is pervaded by these ideas, and that scholar was full of praise for the way in which D'Alembert had applied his principles to explain the precession of the equinoxes.

Among D'Alembert's contemporaries, there was one who read the *Treatise* from beginning to end, studying it with pencil in hand: that was Bézout, and all because D'Alembert had asked him to annotate the second edition (1758) to make it more accessible to a wider readership. A careful examination of Bézout's 61 notes certainly confirms that he had read and understood the *Treatise* in depth. Were there any other scholars in the second half of the Enlightenment who could say the same? One must seriously doubt it: the difficulty of the book, the time at their disposal and the disparity of their specializations reduce the number of candidates, although it may be that Bossut, a close follower of D'Alembert and the author of textbooks on mechanics for gifted students, looked through the *Treatise* with some seriousness.

There is no doubt that the book has had a considerable impact on mathematics and theoretical mechanics, at least in the long term. But this has nearly always been indirect, as 19th-century authors looked on D'Alembert's book merely as a stepping stone to Lagrange's *Méchanique analitique*. While these scholars, from W.R. Hamilton to Ernst Mach, had certainly glanced at the original, it is often very difficult to distinguish between direct and indirect knowledge in their writings. It must be remembered that, from the viewpoint of the 19th century, a large part of the book is difficult to understand because of the notation, the rather archaic style, and the idiosyncrasy of D'Alembert's exposition.

There are to our knowledge two annotated editions, both of which have been translated. The first was by Arthur Korn (1870–1945); because of his original ideas on the causes of gravitation and the pre-eminence of the communication of motion in mechanics, as well as the debates among German scholars at the turn of the century, Korn had personal reasons for his interest in D'Alembert. An examination of his annotations shows that Korn had understood the *Treatise* 'in his way', strongly marked by the 19th century; and the result is very different from that which would be required of a modern critical and impersonal edition. The other annotated translation appeared in the USSR in 1950.

Of course, the *Treatise* has also attracted the attention of historians of science. Important aspects have been studied by R. Dugas, C. Truesdell, T. Hankins, C. Fraser, M. Paty, A. Firode and V. Le Ru among others, and one can say that the net result of these researches provides us with a less biased view of the book today. The edition of the complete works now being prepared will offer to a wider readership not only a simultaneous view of both editions (1743 and 1758), the preliminary drafts and a record of the immediate reception of the book, but also the benefit of the explanations needed to decipher its more difficult Sections.

BIBLIOGRAPHY

Dhombres, J. et Radelet, P. 1991 'Contingence et nécessité en mécanique. Etude de deux textes inédits de Jean D'Alembert', *Physis*, *28*, 35–114.

Dugas, R. 1950. *Histoire de la mécanique*, Paris: Griffon.

Firode, A. 2001. *La dynamique de D'Alembert*, Montréal: Bellarmin; Paris: Vrin.

Fraser, C. 1985. 'D'Alembert's Principle: the original formulation and application in Jean d'Alembert's *Traité de dynamique*', *Centaurus*, *28*, 31–61, 145–159.

Hankins, T. 1970. *Jean d'Alembert. Science and the Enlightenment*, Oxford: Oxford University Press.

Mach, E. 1904. *La mécanique. Exposé historique et critique de son développement* (trad. E. Bertrand), Paris: Hermann.

Le Ru, V. 1994. *Jean le Rond d'Alembert philosophe*, Paris: Vrin.

Passeron, I. et Köllving, U. (eds.) 2002. *Sciences, musique, lumières. hommage à Anne-Marie Chouillet*, Ferney-Voltaire: Centre International d'Etudes sur le Dix-Huitième Siècle. [Contains several articles on D'Alembert and mechanics.]

Paty, M. 1998. *D'Alembert*, Paris: Les Belles Lettres.

Recherches sur Diderot et sur l'Encyclopédie, *21* and *22* (1997). [Contains several articles on D'Alembert and mechanics.]

Viard, J. 2002. 'Le principe de D'Alembert et la conservation du "moment cinétique" d'un système de corps isolés dans le *Traité de dynamique*', *Physis*, *39*, 1–40.

Vilain, C. 2000. 'La question du centre d'oscillation de 1703 à 1743', *Physis*, *37*, 439–466.

CHAPTER 12

LEONHARD EULER, BOOK ON THE CALCULUS OF VARIATIONS (1744)

Craig G. Fraser

In this book Euler extended known methods of the calculus of variations to form and solve differential equations for the general problem of optimizing single-integral variational quantities. He also showed how these equations could be used to represent the positions of equilibrium of elastic and flexible lines, and formulated the first rigorous dynamical variational principle.

First publication. Methodus inveniendi lineas curvas maximi minimive proprietate gaudentes, sive solutio problematis isoperimetrici latissimo sensu accepti, Lausanne and Geneva: Bousquet, 1744. 320 pages.

Later edition. As Euler, *Opera omnia*, series 1, vol. 24 (ed. Constantin Carathéodory), Zurich: Orell Fussli, 1952.

Partial German translations. 1) Chs. 1, 2, 5 and 6 in *Abhandlungen über Variationsrechnung* (ed. Paul Stäckel), Leipzig: Engelsmann, 1894 (*Ostwald's Klassiker der exakten Wissenschaften*, no. 46). 2) App. 1 in *Abhandlungen über das Gleichgewicht* ... (ed. H. Linsenbarth), Leipzig: Engelsmann, 1910 (*Ostwald's Klassiker*, no. 175).

Related articles: Newton (§5), Leibniz, Euler and Lagrange on the calculus (§4, §14, §19).

1 INTRODUCTION

Euler's *Methodus inveniendi* was the first of a series of books that he wrote on calculus in the 1740s and the years that followed; notable later works were the *Introductio* of 1748 on infinite series and the *Institutiones* of 1755 and 1768–1774 on the differential and integral calculus (§13, §14). Although the *Methodus inveniendi* was published in 1744, it was completed by 1741, and was written when Euler was a young man in his late twenties and early thirties at the Academy of Sciences in Saint Petersburg. Born in 1707 to a pastor in Basel in

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

Switzerland, he had quickly showed his mathematical abilities, especially under the tutelage of Johann Bernoulli (1667–1748). His career fell into three parts, all served under some kind of monarchical support. The first and third parts were passed at the (new) Academy in Saint Petersburg: from 1727 to 1741 (when he wrote the *Methodus inveniendi*), and from 1766 to his death in 1783. In between he worked at the Academy in Berlin, where he wrote the other two writings that feature in this book. Apart from this trio, he was extraordinarily prolific, contributing importantly to virtually all areas of mathematics of his day [Thiele, 1982].

The *Methodus inveniendi* is of two-fold interest for historians of mathematics. First, it was a highly successful synthesis of what was then known about problems of optimization in the calculus, and presented general equational forms that became standard in the calculus of variations. Euler's method was taken up by Joseph Louis Lagrange (1736–1813) 20 years later and brilliantly adapted to produce a novel technique for solving variational problems (§16). The two appendices to Euler's book applied variational ideas to problems in statics and dynamics, and these too became the basis for Lagrange's later researches. Second, in Euler's book some of his distinctive contributions to analysis appear for the first time or very nearly the first time: the function concept, the definition of higher-order derivatives as differential coefficients; and the recognition that the calculus is fundamentally about abstract relations between variable quantities, and only secondarily about geometrical curves. The *Methodus inveniendi* is an important statement of Euler's mathematical philosophy as it had matured in the formative years of the 1730s.

2 ORIGINS AND BASIC RESULTS

The early Leibnizian calculus consisted of a sort of geometrical analysis in which differential algebra was employed in the study of 'fine' geometry (§4.2). The curve was analysed in the infinitesimal neighbourhood of a point and related by means of an equation to its overall shape and behaviour. An important curve that was the solution of several variational problems was the cycloid, the path traced by a point on the perimeter of a circle as it rolls without slipping on a straight line. This curve appeared on the frontispiece of Euler's *Methodus inveniendi* (Figure 1) and was a kind of icon of the early calculus. The cycloid possessed a simple description in terms of the infinitesimal calculus. Let the generating circle of radius r roll along the x-axis and let the vertical distance be measured downward from the origin along the y-axis (Figure 2). An elementary geometrical argument revealed that the equation of the cycloid is

$$\left(\frac{ds}{dy}\right)^2 = \frac{2r}{y}, \tag{1}$$

where $ds = \sqrt{(dx^2 + dy^2)}$ is the differential element of path length.

The cycloid was most notably the solution to the brachistochrone problem. Consider a curve joining two points in a vertical plane and consider a particle constrained to descend along this curve. It is necessary to find the curve for which the time of descent is a minimum. Let us take the origin as the first point and let the coordinates of the second be $x = a$ and $y = b$. We assume the particle begins from rest. By Galileo's law the speed of a particle

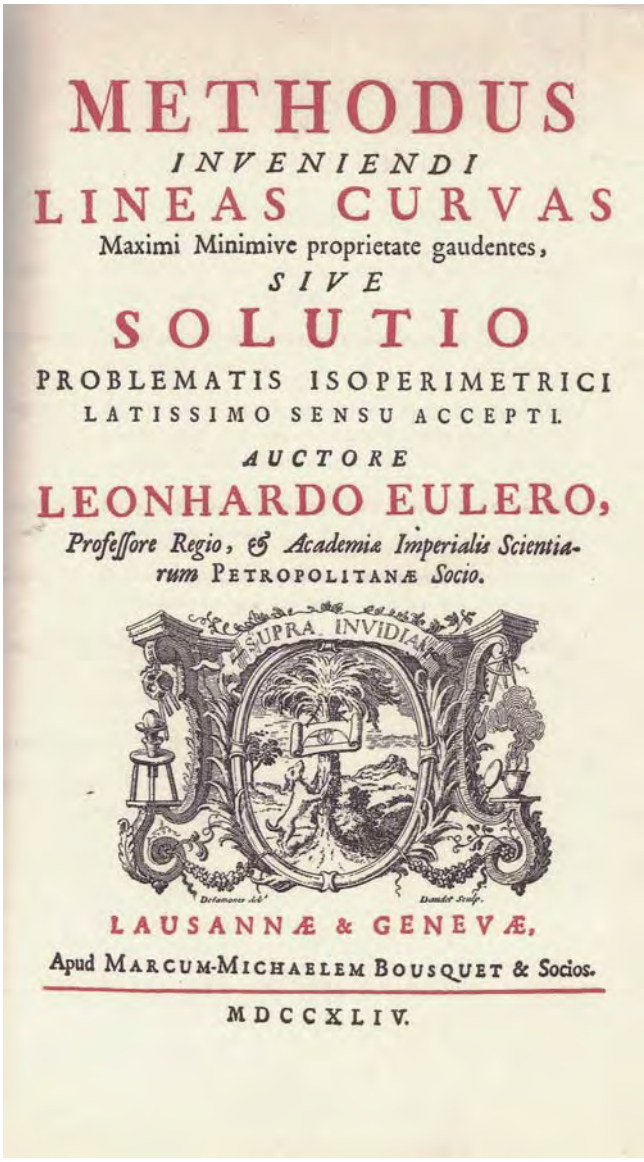

METHODUS
INVENIENDI
LINEAS CURVAS
Maximi Minimive proprietate gaudentes,
SIVE
SOLUTIO
PROBLEMATIS ISOPERIMETRICI
LATISSIMO SENSU ACCEPTI.
AUCTORE
LEONHARDO EULERO,
Professore Regio, & Academiæ Imperialis Scientiarum PETROPOLITANÆ Socio.
SUPRA INVIDIAM
LAUSANNÆ & GENEVÆ,
Apud MARCUM-MICHAELEM BOUSQUET & Socios.
MDCCXLIV.

Figure 1.

in constrained fall when it has fallen a distance y is $\sqrt{(2gy)}$, where g is an accelerative constant. We have the relations

$$\frac{ds}{dt} = \sqrt{2gy} \quad \text{or} \quad dt = \frac{1}{\sqrt{2gy}}\,ds = \frac{\sqrt{1+y'^2}\,dx}{\sqrt{2gy}}. \tag{2}$$

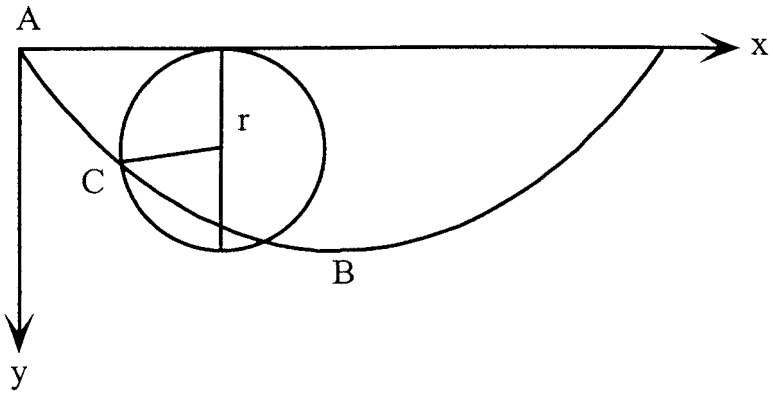


Figure 2.

Hence the total time of descent is given by the integral

$$T = \frac{1}{\sqrt{2g}} \int_0^a \frac{\sqrt{1+y'^2}}{\sqrt{y}}\, dx. \tag{3}$$

The problem of the brachistochrone is to find the particular curve $y = y(x)$ that minimizes this integral.

Following Johann Bernoulli's public challenge in 1696 solutions to this problem were devised by his elder brother Jakob, by Johann himself and by Isaac Newton and G.W. Leibniz. They all showed that the condition that the time of descent is a minimum leads to (1) and, with the exception of Leibniz, concluded that the given curve is a cycloid. Johann's solution was based on an optical–mechanical analogy that is well-known today from its description by Ernst Mach in his *Die Mechanik in ihrer Entwicklung historisch-kritisch dargestellt* (1883). Although of interest, his solution did not provide a suitable basis for further work in the subject.

Jakob Bernoulli's solution on the other hand was illustrative of the ideas that would develop into the calculus of variations. He considered any three points C, G and D on the hypothetical minimizing curve, where the points are assumed to be infinitesimally close to each other. He constructed a second neighbouring curve identical to the first except that the arc *CGD* was replaced by *CLD* (Figure 3). Because the curve minimizes the time of descent it is clear that the time to traverse *CGD* is equal to time to traverse *CLD*. Using this condition and the dynamical relation $ds/dt \propto \sqrt{y}$ Bernoulli was able to derive (1).

Jakob Bernoulli also investigated problems in which the minimizing or maximizing curve satisfied an auxiliary integral condition. The classical isoperimetric problem was the prototype for this class of examples. His idea was to vary the curve at two successive ordinates, thereby obtaining an additional degree of freedom, and use the side constraint to derive a differential equation. Although Jakob died in 1705, some of his ideas were taken up by Brook Taylor in his *Methodus incrementorum* of 1715. Taylor skillfully developed and refined Jakob's conception, introducing some important analytical innovations of his own. Stimulated by Taylor's research, and concerned to establish his brother's priority,

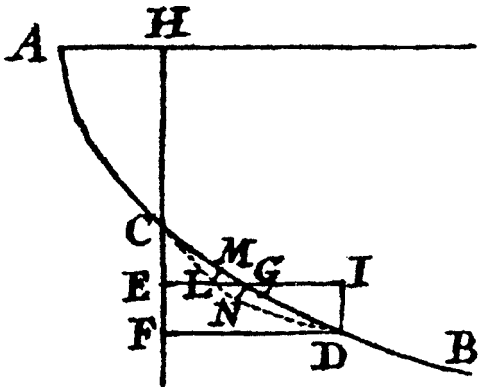


Figure 3.

Johann, then thirty-eight, also adopted Jakob's methods and developed them along more geometric lines in a paper that was published in 1719.

In two memoirs published in the St. Petersburg Academy of Sciences in 1738 and 1741, Euler extracted from the various solutions of Jakob and Johann Bernoulli, as well as the researches of Taylor, a general approach to single-integral variational problems. These investigations were further developed and became the subject of the *Methodus inveniendi*, of which the contents is summarised in Table 1. Its title may be translated 'The method of finding plane curves that show some property of maximum or minimum, or the solution of isoperimetric problems in the widest accepted sense'.

Euler realized that the different integrals in the earlier problems were all instances of the single form

$$\int_a^b Z\left(x, y, y', \ldots, y^{(n)}\right) dx, \tag{4}$$

where Z is a function of x, y and the first n derivatives of y with respect to x. He derived a differential equation, known today as the Euler or Euler–Lagrange equation, as a fundamental condition that must be satisfied by a solution of the variational problem.

Table 1. Contents by Chapters of Euler's book.

Part	**Page**	**Content**
Ch. 1	1	'Method of maximum and minimum' in general.
Ch. 2	32	Differential equations for the optimizing curve.
Ch. 3	83	Side conditions in the form of differential equations.
Ch. 4	130	Resolution of various problems.
Chs. 5–6	171	Isoperimetric problems.
App. 1	245	Elastic curves.
App. 2	311	Principle of least action. [End 320.]

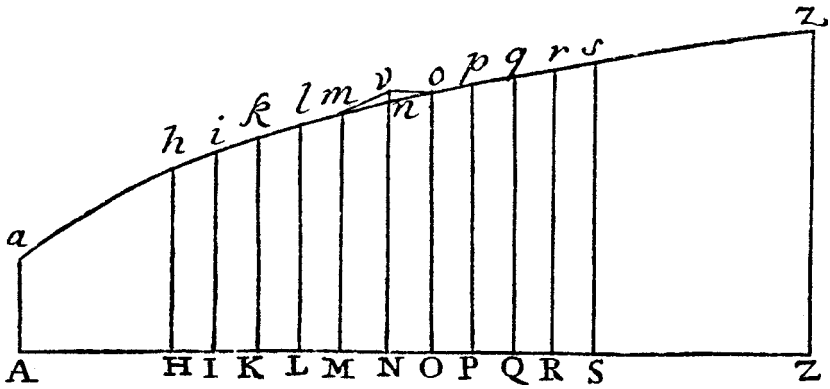


Figure 4.

In Chapter 2 Euler developed his derivation of this equation (for the case $n = 1$) with reference to Figure 4, in which the line *anz* is the hypothetical extremizing curve. The letters M, N, O designate three points of the x-axis AZ infinitely close together. The letters m, n, o designate corresponding points on the curve given by the ordinates Mm, Nn, Oo. Let $AM = x$, $AN = x'$, $AO = x''$ and $Mm = y$, $Nn = y'$, $Oo = y'$. The differential coefficient p is defined by the relation $dy = p\,dx$; hence $p = dy/dx$. We have the following relations

$$p = \frac{y' - y}{dx}, \qquad p' = \frac{y'' - y'}{dx}. \tag{5}$$

The integral $\int_a^b Z\,dx$ was regarded by Euler as an infinite sum of the form $\cdots + Z$, $dx + Z\,dx + Z'\,dx + \cdots$, where Z, is the value of Z at $x - dx$, Z its value at x and Z' its value at $x + dx$, and where the summation begins at $x = a$ and ends at $x = b$. It is important to note that Euler did not employ limiting processes or finite approximations. Let us increase the ordinate y' by the infinitesimal 'particle' nv, obtaining in this way a comparison curve *amvoz*. Consider the value of $\int_a^b Z\,dx$ along this curve. By hypothesis the difference between this value and the value of $\int_a^b Z\,dx$ along the actual curve will be zero. The only part of the integral that is affected by varying y' is $Z\,dx + Z'\,dx = (Z + Z')\,dx$. Euler wrote:

$$dZ = M\,dx + N\,dy + P\,dp, \qquad dZ' = M'\,dx + N'\,dy' + P'\,dp'. \tag{6}$$

He proceeded to interpret the differentials in (6) as the infinitesimal changes in Z, Z', x, y, y', p, p' that result when y' is increased by nv. From (5) we see that dp and dp' equal nv/dx and $-nv/dx$. (These changes were presented by Euler in the form of a table, with the variables in the left column and their corresponding increments in the right column.) Hence (6) becomes

$$dZ = P \cdot \frac{nv}{dx}, \qquad dZ' = N' \cdot nv - P' \cdot \frac{nv}{dx}. \tag{7}$$

Thus the total change in $\int_a^b Z\,dx$ equals $(dZ + dZ')\,dx$ or $nv \cdot (P + N'\,dx - P')$. This expression must be equated to zero. Euler set $P' - P = dP$ and replaced N' by N. He therefore obtained $0 = N\,dx - dP$ or

$$N - \frac{dP}{dx} = 0, \tag{8}$$

as the final equation of the problem.

Equation (8) is the simplest instance of the Euler differential equation, giving a condition that must be satisfied by the minimizing or maximizing arc. Noting that N and P are the partial derivatives of Z with respect to y and y' respectively, we may write (8) in modern notation as

$$\frac{\partial Z}{\partial y} - \frac{d}{dx}\frac{\partial Z}{\partial y'} = 0. \tag{9}$$

He also derived the corresponding equation when higher-order derivatives of y with respect to x appear in the variational integral. This derivation was a major theoretical achievement, representing the synthesis in one equational form of the many special cases and examples that had appeared in the work of earlier researchers.

3 FOUNDATIONS OF ANALYSIS

Near the beginning of his book Euler noted that a purely analytical interpretation of the theory is possible. Instead of seeking the curve which makes W an extremum one seeks that 'equation' between x and y which among all such equations when introduced into (1) makes the quantity W a maximum or minimum (p. 13). He wrote:

> Corollary 8. In this way questions in the doctrine of curved lines may be referred back to pure analysis. Conversely, if questions of this type in pure analysis be proposed, they may be referred to and solved by means of the doctrine of curved lines.
> Scholium 2. Although questions of this kind may be reduced to pure analysis, nevertheless it is useful to consider them as part of the doctrine of curved lines. For though indeed we may abstract from curved lines and consider absolute quantities alone, so these questions at once become abstruse and inelegant and appear to us less useful and worthwhile. For indeed methods of resolving these sorts of questions, if they are formulated in terms of abstract quantities alone, are very abstruse and troublesome, just as they become wonderfully practical and become simple to the understanding by the inspection of figures and the linear representation of quantities. So although questions of this kind may be referred to either abstract or concrete quantities it is most convenient to formulate and solve them by means of curved lines. Thus if a formula composed of x and y is given, and that equation between x and y is sought such that, the expression for y in terms of x given by the equation being substituted, there is a maximum or minimum; then we can always transform this question to the determination of the curved line, whose abscissa is x and ordinate is y, for

> which the formula W is a maximum or minimum, if the abscissa x is assumed to have a given magnitude.

Euler's view seems to have been that while it is possible in principle to approach the calculus of variations purely analytically it is more effective in practice to refer problems to the study of curves. This conclusion could hardly have seemed surprising. Each of the various examples and problems which historically made up the subject had as its explicit goal the determination of a curve; the selection of such objects was part of the defining character of this part of mathematics. What is perhaps noteworthy about Euler's discussion is that he should have considered the possibility at all of a purely analytical treatment.

The basic variational problem of maximizing or minimizing (4) involves the selection of a curve from among a class of curves. In the derivation of (8) the variables x and y are regarded as the orthogonal Cartesian coordinates of a curve. Each of the steps in this derivation involves reference to the geometrical diagram in Figure 4 above. In Chapter 4, however, Euler returned to the point of view that he had indicated at the beginning of the treatise. In the opening proposition the variational problem is formulated as one of determining that 'equation' connecting two variables x and y for which a magnitude of the form (4) (given for the general case where higher-order derivatives and auxiliary quantities are contained in Z) is a maximum or minimum. In his solution he noted that such variables can always be regarded as orthogonal coordinates and so determine a curve. The solution then follows from the theory developed in the preceding chapters. In the first corollary he wrote:

> Thus the method presented earlier may be applied widely to the determination of equations between the coordinates of a curve which makes any given expression $\int Z\,dx$ a maximum or a minimum. Indeed it may be extended to any two variables, whether they involve an arbitrary curve, or are considered purely in analytical abstraction.

Euler illustrated this claim by solving several examples using variables other than the usual rectangular Cartesian coordinates. In the first example he employed polar coordinates to find the curve of shortest length between two points (Figure 5). We are given the points A and M and a centre C; it is necessary to find the shortest curve AM joining A and M. Let x be the pole angle ACM and y the radius CM. Because the differential element of path-length is equal to $\sqrt{(dy^2 + y^2\,dx^2)}$ the formula for the total path-length is $\int dx\,\sqrt{(yy + pp)}$, where $p\,dx = dy$ and the integral is taken from $x = 0$ to $x = \angle ACM$. Here x does not appear in the integrand Z of the variational integral, so that $dZ = N\,dy + P\,dp$. The equation (8) gives $N = dP/dx$ so that we have $dZ = dP\,p + P\,dp$ and a first integral is $Z + C = Pp$, where C is a constant. Since $Z = \sqrt{(yy + pp)}$ we have

$$C + \sqrt{(yy + pp)} = \frac{pp}{\sqrt{(yy + pp)}}, \quad \text{i.e.,} \quad \frac{yy}{\sqrt{(yy + pp)}} = Const. = b. \tag{10}$$

Let PM be the tangent to the curve at M and CP the perpendicular from C to this tangent. By comparing similar triangles in Figure 5 we see that $Mm : Mn = MC : CP$. Since $Mm = dx/\sqrt{y^2 + p^2}$, $Mn = y\,dx$ and $MC = y$ it follows that $CP = y^2/\sqrt{y^2 + p^2}$. Hence CP is a constant. Euler concluded from this property that the given curve AM is a straight line.

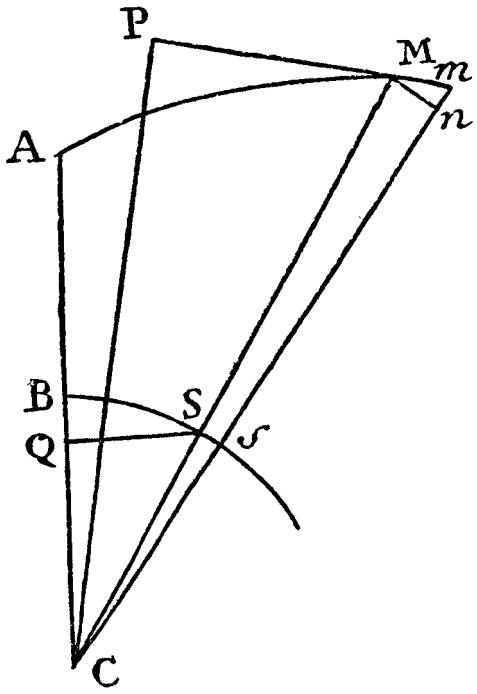


Figure 5.

In the second example Euler displayed a further level of abstraction in his choice of variables. Here we are given the axis *AC* with the points *A* and *P*, the perpendicular line *PM* and a curve *ABM* joining *A* and *M* (Figure 6). Given that the area *ABMP* is some given constant value we must find that curve *ABM* which is of the shortest length. Euler set the

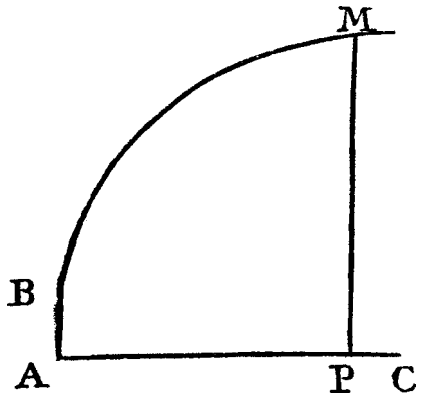


Figure 6.

abscissa $AP = t$, the ordinate $PM = y$ and let x equal the area under the curve from A to P. We have $dx = y\,dt$ and the variational integral becomes $\int \sqrt{(dy^2 + dx^2/yy)}\,dx$. Because x does not appear in the integrand we obtain as before the first integral $Z = C + pP$. Substituting the expressions for Z and P into this integral we obtain

$$\frac{\sqrt{(1 + yypp)}}{y} = C + \frac{ypp}{\sqrt{(1 + yypp)}}. \tag{11}$$

Letting $dx = y\,dt$, we obtain after some further reductions the final equation $t = c \pm \sqrt{(bb - yy)}$. Hence the desired curve is the arc of a circle with its centre on the axis *AP* at the foot P of the ordinate corresponding to M.

A range of non-Cartesian coordinate systems had been employed in earlier mathematics but never with the same theoretical import as in Euler's variational analysis. Here one had a fully developed mathematical process, centred on the consideration of a given analytically-expressed magnitude, in which a general equational form was seen to be valid independent of the geometric interpretation given to the variables of the problem. Thus it is not at all essential in the reasoning employed in the derivation of (9) that the line *AZ* be perpendicular to *Mm* (Figure 4); indeed it is clear that the variable x need not be a length nor even a co-ordinate variable in the usual sense. As Euler observed in the first corollary, the variables of the problem are abstract quantities, and Figure 4 is simply a convenient geometrical visualization of an underlying analytical process.

Euler and later 18th-century analysts broke with the geometrical tradition, but they did not thereby adopt the point of view of modern real analysis. Euler's understanding was very different from our outlook today, in which the expression Z that is to be optimized is any quantity whatsoever formulated in terms of the function $y = y(x)$ and its derivatives. For Euler, the quantities and relations of analysis are always 'given': they arise from definite problems in geometry, mechanics or some other area of mathematical science. He developed an abstract interpretation of the variational formalism—the fundamental objects of study were relations between variables 'given in analytical abstraction'—but his point of view was structured as well by tacit assumptions concerning the logical status of the problems of the subject as things that were given from without. The notion that at the outset one could consider any expression Z defined according to logically prior and autonomous criteria was quite beyond Euler's conceptual horizons and was foreign to the outlook of 18th-century analysis.

4 LATER DEVELOPMENTS: LAGRANGE, EULER AND THE CALCULUS OF VARIATIONS

In his book Euler had noted the somewhat complicated character of his variational process and called for the development of a simpler method or algorithm to obtain the variational equations. Lagrange's first important contribution to mathematics, carried out when he was 19 years old, consisted of his invention of the δ-algorithm to solve the problems of Euler's *Methodus inveniendi.* He announced his new method in a letter of 1755 to Euler, and published it as [Lagrange, 1762] in the Proceedings of the Turin Society. His algorithm permitted the systematic derivation of the variational equations and facilitated the treatment

of conditions at the endpoints. His innovation was immediately adopted by Euler, who introduced the name 'calculus of variations' to describe the subject founded on the new method. Lagrange's new approach originated in his (tacit) recognition that the symbol d was being used in two distinct ways in Euler's derivation of (8). In (8) and the final step by which it is obtained, d was used to denote the differential as it was customarily used and understand in Continental analysis of the period. The differential dx was held constant; the differential of any other variable equalled the difference of its value at x and its value at an abscissa a distance dx from x. By contrast, the differentials dx, dy, etc. that appear in (6) were interpreted by Euler as the changes in x, y, etc. that result when the single ordinate y is increased by the 'particle' nv. Thus the 'differentials' $dy'\,dp$, dp' equal nv, nv/dx, $-nv/dx$; the 'differentials' dx, dy, dp'', etc. are zero.

The young Lagrange had the perspicacity to recognize this dual usage and invented the symbol 'δ' to denote the second type of differential change. Using it he devised a new analytical process to investigate problems of maxima and minima. Although the purpose of his method was to compare curves in the plane, it was nonetheless introduced in a very formal manner. The symbol δ has properties analogous to the usual d of the differential calculus. Thus $\delta(x+y)=\delta x+\delta y$ and $\delta(xy)=x\delta y+y\delta x$. In addition, d and δ are interchangeable ($d\delta=\delta d$) as are d and the integral operation $\int$.

The δ-process led to a new and very simple derivation of the Euler equation (8). It is necessary to determine $y=y(x)$ so that

$$\delta\int_a^b Z\,dx=0, \tag{12}$$

where $Z=Z(x,y,p)$ and $p=dy/dx$. Applying the δ operation to the expression Z we obtain

$$\delta Z=N\delta y+P\delta p. \tag{13}$$

Note that here all of the ordinates are simultaneously being varied, and not just one, as had been the case in Euler's analysis. Because the δ and $\int$ are interchangeable we have

$$\delta\int_a^b Z\,dx=\int_a^b \delta Z\,dx=\int_a^b (N\delta y+P\delta p)\,dx \tag{14}$$

and also $\delta p=\delta(dy/dx)=d(\delta y)/dx$. An integration by parts gives rise to the identity

$$\int_a^b P\delta p\,dx=\int_a^b P\frac{d(\delta y)}{dx}\,dx=P\delta y|_a^b-\int_a^b \frac{dP}{dx}\delta y\,dx. \tag{15}$$

Hence the condition $\delta\int_a^b Z\,dx=0$ becomes

$$P\delta y|_a^b-\int_a^b\left(N-\frac{dP}{dx}\right)\delta y\,dx=0. \tag{16}$$

We suppose that δy is zero at the end values $x = a, b$. (16) then reduces to

$$\int_a^b \left(N - \frac{dP}{dx} \right) \delta y \, dx = 0. \tag{17}$$

From (17) we are able to infer the Euler equation

$$N - \frac{dP}{dx} = 0. \tag{18}$$

Euler took up Lagrange's new method in his writings of the 1760s and 1770s. In a paper published in 1772 he presented what would become the standard interpretation of the δ-process as a means for comparing classes of curves or functions. We assume that y is a function of x and a parameter t, $y = y(x, t)$, where the given curve $y = y(x)$ is given by the value of $y(x, t)$ at $t = 0$. We define δy to be $\frac{\partial y}{\partial t}|_{t=0}\, dt$. (It would be logically more consistent to define $\delta y = \frac{\partial y}{\partial t}|_{t=0} t$, and require that t be small. Euler apparently used dt rather than t so as to indicate explicitly that the multiplicative factor is small.) One way of doing this, Euler explained, is to set $y(x, t) = X(x) + tV(x)$, where $y(x) = X(x)$ is the given curve and $V(x)$ is a comparison or increment function; hence we have $\delta y = dt\, V(x)$. In this conception the variation of a more complicated expression made up of $y(x, t)$ and its derivatives with respect to x is obtained by taking the partial derivative with respect to t, setting $t = 0$ and introducing the multiplicative factor dt. In later variational mathematics the parameter 'ε' would often be used instead of 't'.

BIBLIOGRAPHY

Bernoulli, Jakob. 1697. 'Solutio problematum fraternorum, peculiari programmate cal. Jan. 1697 Groningae, nec non Actorum Lips. mens Jun. & Dec. 1696, & Febr. 1697 propositorum: una cum propositione reciproca aliorum', *Acta eruditorum*, 211–217. [Repr. in *Opera* (1744), 768–778; and in [Bernoulli, 1991], 271–282.]

Bernoulli, Jakob. 1701. 'Analysis magni problematis isoperimetrici', *Acta eruditorum*, 213–228. [Repr. in *Opera* (1744), 895–920; Johann's *Opera*, vol. 2 (1742), 219–234; and in [Bernoulli, 1991], 485–505.]

Bernoulli, Jakob. 1744. *Jacobi Bernoulli Basileensis opera* (ed. Gabriel Cramer), 2 vols., Lausanne and Geneva: G. and P. Cramer.

Bernoulli, Johann. 1719. 'Remarques sur ce qu'on a donné jusqu'ici de solutions des Problêmes sur les Isoperimetres, avec une nouvelle méthode courte & facile de les résoudre sans calcul, laquelle s'étend aussi à d'autres problêmes qui ont rapport à ceux-là', *Mémoires de l'Académie Royale des Sciences* (1718), 100–138. [Repr. in [Bernoulli, 1991], 527–568.]

Bernoulli, Johann. 1742. *Johannis Bernoulli opera omnia* (ed. Gabriel Cramer), 4 vols., Lausanne and Geneva: Bousquet. [Repr. Hildesheim: Olms, 1968.]

Bernoulli, Jakob and Bernoulli, Johann. 1991. *Die Streitschriften von Jacob und Johann Bernoulli Variationsrechnung* (ed. H.H. Goldstine with P. Radelet-de-Grave), Basel: Birkhäuser. [With introduction by Goldstine on pp. 4–113.]

Carathéodory, C. 1952. 'Einführung in Eulers Arbeiten über Variationsrechnung', in Euler's *Opera Omnia*, ser. 1, vol. 24, viii–li.

Euler, L. 1738. ‘Problematis isoperimetrici in latissimo sensu accepti solutio generalis’, *Commentarii academiae scientiarum Petropolitanae, 6* (1732–1733), 123–155. [Repr. in Euler *Opera omnia*, ser. 1, vol. 25 (1952), 13–40.]

Euler, L. 1741. ‘Curvarum maximi minimive proprietate gaudentium inventio nova et facilis’, *Ibidem, 8* (1736), 159–190. [Repr. in *Opera omnia*, ser. 1, vol. 25 (1952), 54–80.]

Euler, L. 1772. ‘Methodus nova et facilis calculum variationum tractandi’, *Ibidem, 16* (1771), 35–70. [Repr. in *Opera omnia*, ser. 1, vol. 25 (1952), 208–235.]

Feigenbaum, L. 1985. ‘Brook Taylor and the method of increments’, *Archive for history of exact sciences, 34*, 1–140.

Fraser, C. 1994. ‘The origins of Euler’s variational calculus’, *Archive for history of exact sciences, 47*, 103–141.

Fraser, C. 1996. ‘The background to and early emergence of Euler’s analysis’, in M. Otte and M. Panza (eds.), *Analysis and synthesis in mathematics. History and philosophy*, Dordrecht: Kluwer, 47–78.

Goldstine, H.H. 1980. *A history of the calculus of variations from the 17th through the 19th century*, New York: Springer.

Lagrange, J.L. 1762. ‘Essai sur une nouvelle méthode pour déterminer les maxima et les mimima des formules intégrales indéfinies’, *Miscellanea physico-mathematica societatis Taurinensia, 2* (1760–1761), 407–418. [Repr. in *Oeuvres*, vol. 1, 333–362.]

Taylor, B. 1715. *Methodus incrementorum directa & inversa*, London: Innys.

Thiele, R. 1982. *Leonhard Euler*, Leipzig: Teubner.

CHAPTER 13

LEONHARD EULER, 'INTRODUCTION' TO ANALYSIS (1748)

Karin Reich

In the first volume of this book Euler brought together old and new results on functions and (infinite) series, and related notions such as continued fractions. In the second volume he handled aspects of analytic, coordinate and differential geometry. The book was a basic source of approaches and information in these areas for a long time.

First publication. Introductio in analysin infinitorum, 2 volumes, Lausanne: Marcus-Michaelis Bousquet & Socii, 1748. xvi + 320 pages; 398 pages and 40 plates.

Photoreprint. Brussels: Culture et Civilisation, 1967.

Reprints. 1) Lausanne: J.H. Pott, 1783. 2) As *Opera omnia*, ser. 1, vol. 8 (ed. Adolf Krazer and Ferdinand Rudio) and vol. 9 (ed. Andreas Speiser), Leipzig und Berlin: Teubner, 1922 and 1945.

New edition. 2 volumes. Lyons: Bernuset, Delamolliere, Falque & Soc., 1797.

French translation of volume 1. Introduction à l'analyse des infiniments petits (trans. M. Pezzi), Strasbourg: Librairie Académique, 1786.

Full French translation. Introduction à l'analyse infinitésimale (trans. J.B. Labey), 2 vols., Paris: Barrois, 1796–1797. [Repr. Paris: Bachelier, 1835.]

Full German translation. Einleitung in die Analysis des Unendlichen (trans. Johann Andreas Christian Michelsen), 2 vols., Berlin: Carl Matzdorff, 1788. [Repr. Berlin: Reimer, 1835–1836. Photorepr. Berlin: Springer, 1983.]

German translation of volume 1. Einleitung in die Analysis des Unendlichen (trans. Hermann Maser), Berlin: Springer, 1885. [Photorepr. Berlin: Springer, 1983.]

English translation. Introduction to the analysis of the infinite (trans. John D. Blanton), 2 vols., New York: Springer, 1988, 1990.

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

Spanish translation. *Introducción análisis de los infinitos* (trans. José Luis Arantegui Tamayo, notes by Antonio José Duran Guardeno), 2 vols., Seville: Sociedad Andaluza de Educación Matematica 'Thales', 2000.

Russian translations. 1) *Vvedenie v analiz beskonechno malych*, 2 vols., Moscow and Leningrad: 1936. 2) *Vvedenie v analiz beskonechnych* (trans. V.S. Gochman, ed. I.B. Pogrebysky), 2 vols., Moscow: Gosudarstvennoe Izdatelstvo Fiziko-Matematiceskoi Literatury, 1961.

Related articles: Leibniz (§4), Euler (§12, §14), Lagrange (§19), Monge (§17), Lacroix (§20), Cauchy on real-variable analysis (§25).

1 INTRODUCTION TO THE *INTRODUCTIO*

Euler's book is one of the few books on mathematics that is mentioned by John Carter and Percy H. Muir in their *Printing and the mind of Man* (1967). There Euler is compared with Euclid: what Euclid did with his *Elements* for geometry, Euler did with his *Introductio* for analysis. It was by means of this textbook that analysis became an independent discipline within mathematics.

Euler's biography was noted in §12.1; see also [Fellmann, 1983] and [Thiele, 2000]. During his first stay in Saint Petersburg, up to 1741, he had published many articles and some books, such as the *Mechanica* [Euler, 1736] and *Tentamen novae theoriae musicae* (1739). In 1741 he moved to Berlin, where he became professor of mathematics at the Prussian Academy; the head was Friedrich II (1712–1786), who had reigned since 1740. In the same year Pierre-Louis Moreau de Maupertuis (1698–1759) came to Berlin, where he became President of the Academy in 1746. During the years 1744–1766 Euler was appointed 'Director of the Mathematical Class'.

The *Introductio* was not Euler's first purely mathematical book, for in 1744 he had published his *Methodus inveniendi lineas curvas*, described in §12. Apparently that year he also completed the new book, which was its successor. It was also the first part of an 'analytical trilogy' that was to appear in Saint Petersburg: *Introductio*, *Institutiones calculi differentialis* (1755: §14) and *Institutiones calculi integralis* (3 volumes, 1768–1770).

With his *Introductio* Euler wanted to present a textbook containing all the topics necessary to know before beginning to study the infinitesimal calculus. It is divided into two 'Books'. Its contents are summarised in Table 1; for a general survey see [Cantor, 1898, chs. 111 and 115]. The title page is shown in Figure 1.

2 BOOK I, ON ANALYSIS

2.1 The significance of 'function'. The first Book is explained on its own title page as 'containing an explanation of functions of variable quantities; the resolution of functions into factors and their development in infinite series; together with the theory of logarithms, circular arcs, and their sines and tangents, also many other things which are no little aid in the study of analysis'. While not the first textbook on analysis, Euler's *Introductio* was the first to take the concept of function as its ground: 'A function of a variable quantity

Table 1. Contents by chapters of Euler's book.
All titles are translated. Volume I has its own numeration of chapters, pages and articles; volume 2 has two sequences for chapters and articles, one for curves and surfaces respectively. The edition in the *Opera omnia* indicates the original pagination, which is used here.

Chap.	**Page**	**Art.**	**Title**
	ii–xiii		Dedication, preface.
I	3	1	On functions in general.
II	15	27	On the transformation of functions.
III	36	46	On the transformation of functions by substitution.
IV	46	59	On the expansion of functions in infinite series.
V	60	77	On functions of two or several variables.
VI	69	96	On exponential quantities and logarithms.
VII	85	114	On the expansion of exponential quantities and of logarithms in series.
VIII	93	126	On transcendental quantities arising from the circle.
IX	107	143	On the investigation of trinomial factors.
X	128	165	On the use of invented factors in defining the sums of infinite series. [Infinite products.]
XI	145	184	On other infinite expressions for the arc and the sine.
XII	161	199	On the development of real functions of fractions. [Rational functions.]
XIII	175	211	On recurrent series.
XIV	198	234	On the multiplication and division of angles.
XV	221	264	On series arising from the development of factors.
XVI	253	297	On the partition of numbers.
XVII	276	332	On the use of recurrent series in investigating the roots of equations.
XVIII	295	356	On continued fractions. [End 320, art. 382.]
I	3	1	On curved lines in general.
II	12	23	On the change of coordinates.
III	23	47	On the division of algebraic curves into orders.
IV	32	66	On the principal properties of lines of any order.
V	41	85	On lines of the second order.
VI	64	131	On the sub-division of lines of the second order in general.
VII	83	166	On the investigation of branches that extend to the infinite.
VIII	99	198	On asymptotic lines.
IX	114	219	On the sub-division of lines of the third order into types.
X	127	239	On the principal properties of lines of the third order.
XI	139	260	On lines of the fourth order.
XII	150	272	On the exploration of the shapes of curved lines.

Table 1. (*Continued*)

XIII	156	285	On the characteristics of curved lines.
XIV	166	304	On the curvature of curved lines.
XV	181	337	On curves which have one or several diameters.
XVI	194	364	On the determination of curves from given properties of the ordinates.
XVII	212	391	On the determination of curves from other properties.
XVIII	236	435	On the similarity and affinities of curved lines.
XIX	247	457	On the intersection of curves.
XX	269	486	On the construction of equations.
XXI	284	506	On transcendental curved lines.
XXII	304	529	Solution of some problems pertaining to the circle. [End 320, art. 540.]
I	323	1	On the surfaces of bodies in general.
II	337	26	On the sections of surfaces made by whatever planes.
III	348	52	On the sections of cylinders, cones and globes.
IV	365	86	On the interchange of coordinates.
V	373	101	On surfaces of the second order.
VI	388	131	On the intersection of two surfaces.[1] [End 398, art. 152.]

[1]On the contents page (vol. 1, xvi) this title is misstated as 'On the mutual intersection of surfaces'.

is an analytical expression composed in any way whatsoever of the variable quantity and numbers of constant quantities' (art. 4). The roots of this definition of function can be traced back to Johann Bernoulli (1667–1748).

In art. 7 Euler divided the functions into algebraic and transcendental ones, and in art. 8 the algebraic functions into 'rational' and 'irrational' functions. (Sadly, the modern English translator has named the first category 'non-rational'.) Rational functions are 'such that the variable quantity is in no way involved with irrationality; the latter are those in which the variable quantity is affected by radical signs'. Both kinds can be developed into infinite series. Euler even allowed generalized exponents, not only positive integers (art. 59). He then tried to give a full treatment of functions, their transformation as well as their development into infinite series. Some of the summations of divergent series would not now be regarded as correct; but Euler saw himself as finding formal relationships between series and their sum functions, without having the full grasp of summability theory as we now understand it [Ferraro, 1998].

2.2 Exponentials and logarithms. The idea of logarithms came up at the beginning of the 17th century with John Napier and Jost Bürgi. The main idea was to effect a comparison between an arithmetical and a geometric series. So at first the fact $a^0 = 1$ was not put forward, which meant that the relationship between basis, exponent and logarithm was not emphasized.

INTRODUCTIO
IN ANALYSIN
INFINITORUM.

AUCTORE

LEONHARDO EULERO,

Professore Regio BEROLINENSI, *& Academiæ Imperialis Scientiarum* PETROPOLITANÆ *Socio.*

TOMUS PRIMUS.

LAUSANNÆ,

Apud MARCUM-MICHAELEM BOUSQUET & Socios.

MDCCXLVIII.

Figure 1.

Euler was the first who held a clear view of what was needed, in his Chapter 6 'On exponentials and logarithms' (arts. 96–113). In art 102 he introduced the term 'basis'. The exponent z being the variable and a a constant, the expression $a^z = y$ means: 'This value of z, insofar as it is viewed as a function of y, is called the LOGARITHM of y, to be designated by the symbol 'log y'. Thus, according to his theory the logarithm depends upon a constant a, which therefore is called the basis of logarithms. 'Whatever logarithmic base we choose, we always have $\log 1 = 0$' (art. 103). In the following pages Euler explained the operations of multiplication, division and root extraction by means of logarithms. He also mentioned that logarithms normally are transcendental: 'it follows that the logarithm of a number will not be a rational number unless the given number is a power of the base a. Logarithms which are not the powers of the base are neither rational nor irrational, it is with justice that they are called transcendental quantities' (art. 105). The logarithms on the basis of 10 are called the common logarithms; the whole number is the characteristic and the decimal fraction the mantissa (art. 112).

2.3 The derivation of the number 'e' and its importance within logarithms. In the following Chapter 7, 'Exponentials and logarithms expressed through series', Euler went on to show how logarithms could be expressed by infinite series. He was able to derive the 'most natural and fruitful concept of logarithms' by means of the series

$$e = 1 + 1/1 + 1/1 \cdot 2 + 1/1 \cdot 2 \cdot 3 + \cdots = 2 \cdot 7182818284590452353602 8\ldots. \qquad (1)$$

(He had already introduced the letter 'e' as the basis of natural logarithms in a paper of 1728, and used it especially in [Euler, 1744] on continued fractions.) He called logarithms calculated on this basis 'natural or hyperbolic' (art. 122).

2.4 The trigonometric functions. Until Euler the trigonometric quantities were not thought to be functions, but as lines in the circle; but he realised the potential of the functional view. For him sine, cosine, and so on were at first transcendental quantities similar to e but of a different kind; only secondarily were they to be interpreted in terms of arcs of a circle. His main idea was that when imaginary quantities were included these trigonometric functions were dependent on logarithms and exponential quantities. To make calculations easier, Euler initiated a standard circle, the radius of which was 1.

Euler began with the value of π—a symbol that he first used in [1736, art. 287], to represent half the circumference of a circle with the radius 1—which he calculated to 126 decimal places. Using the addition theorems

$$\sin(y + z) = \sin y \cos z + \cos y \sin z \qquad (2)$$

and

$$\cos(y + z) = \cos y \cos z - \sin y \sin z, \qquad (3)$$

together with the equation

$$\sin^2 z + \cos^2 z = 1, \quad \text{or} \quad (\cos z + \sqrt{-1}\sin z)(\cos z - \sqrt{-1}\sin z) = 1, \qquad (4)$$

he derived after a long calculation several results concerning $\pi/4$ and $\pi/6$; they were not new but he had deduced them in a new way. Later he also considered the more complicated series $\pi^2/8$, $\pi^3/32$, $\pi^4/96$, $5\pi^2/1536$, $\pi^6/960, \ldots$. (art. 175). He also put forward the equations

$$\cos v = 1/2\left(e^{v\sqrt{-1}} + e^{-v\sqrt{-1}}\right) \quad \text{and} \quad \sin v = 1/2\sqrt{-1}\left(e^{v\sqrt{-1}} - e^{-v\sqrt{-1}}\right) \tag{5}$$

(art. 138); but, contrary to popular belief, neither here nor anywhere else did he state the now famous result given by $v = \pi$, namely, $e^{\pi\sqrt{-1}} + 1 = 0$. He was to introduce the symbol 'i' for $\sqrt{-1}$ much later, in [Euler, 1794].

2.5 Recurrent series. This kind of series, first introduced by Abraham de Moivre (1667–1754) (§7), plays a major role in the *Introductio*. As Euler pointed out in arts. 62–70, these series arise in rational functions by division: any term is determined by a certain number of the preceding terms, on the basis of a certain fixed law. In Chapter 13 he tried to express any recurrent series by means of simpler recurrent series, and he determined their sums. In Chapter 17 he showed that recurrent series could be useful for the calculation of the roots of an equation.

2.6 Continued fractions. Known to mathematicians before Euler, he had published on them first in [Euler, 1744], where he had used the term 'fractio continua' for the first time. There he had given them the following representation:

$$a + \cfrac{\alpha}{b + \cfrac{\beta}{c + \cfrac{\gamma}{d + \cfrac{\delta}{e + \cdots}}}}. \tag{6}$$

For him they were a third kind of infinite expression, to join infinite series and infinite products. He gave the following definition: 'By a continued fraction I mean a fraction of such a kind that the denominator consists of the sum of an integer and a fraction whose denominator again is the sum of an integer and a fraction of the same kind. This kind of process can continue indefinitely or can stop at some point' (art. 357). They consisted of two kinds of quantities: $a, b, c, \ldots$ as 'denominatores' and $\alpha, \beta, \gamma, \ldots$ as 'numeratores'. He also transformed especially infinite series into continuous fractions, for example (art. 369):

$$\log 2 = 1 - 1/2 + 1/3 - 1/4 + 1/5 - \cdots \tag{7}$$

$$= 1\cfrac{1}{1 + \cfrac{1}{1 + \cfrac{4}{1 + \cfrac{9}{1 + \cfrac{16}{1 + \cfrac{25}{1 + \cdots}}}}}}. \tag{8}$$

Continued fractions have the advantage, that their sums can be approximated quite quickly, and so form a much more convenient way than summing up the successive members of an infinite series. They can also be used for the extraction of roots and for solving quadratic equations (arts. 377–380). Furthermore, every fraction can be developed into a continuous fraction. As an example Euler developed $(e-1)/2$ into

$$\frac{e-1}{2} = \cfrac{1}{1+\cfrac{1}{6+\cfrac{1}{10+\cfrac{1}{14+\cfrac{1}{18+\cfrac{1}{22+\cdots}}}}}}. \tag{9}$$

The case of π yielded the fractions

$$1/0,\ 3/1,\ 22/7,\ 335/106,\ 355/113,\ 103{,}993/33{,}102,\ \ldots, \tag{10}$$

where the third is the approximation of Archimedes and the fifth the approximation of Metius (art. 382).

3 BOOK II ON PLANE AND SURFACE GEOMETRY

The second Book is described as 'containing the theory of curves and an appendix on surfaces'. Though it is not as well known as the first Book, it still contained plenty of ingenious ideas [Boyer, 1956, 179–191]. Euler treated mainly the analytical geometry of the plane, that is, the theory of curves for about 200 pages; his surface theory was 85 pages. He explained in great detail not only rectangular but also polar coordinates.

3.1 Classification of curves. At first, Euler tried to use the same arrangement for his analytic geometry as he had used in his algebraic analysis: he introduced functions that expressed the nature of the curves relating the 'abscissa' x and the 'applicata' y. As in the case of functions, there were continuous curves and discontinuous curves. In geometry especially continuous curves are of interest. Further, he distinguished algebraic and transcendental curves; a transcendental curve is expressed by a transcendental equation (art. 15).

In the case of algebraic equations the degree of the equation is the classifying principle (art. 51). The general linear equation is the expression of a curve of the first order, while the general equations of the second degree are the expressions of second-order lines; these are the conic sections.

In the work following the classification of the curves was determined by means of the asymptotes. In the case of higher orders of the curves the distinction of the different kinds gets more and more complicated. Partly following Isaac Newton's classification, Euler treated 16 different species for third-order lines and distinguished between more than 100 different cases for fourth-order lines.

Euler was also interested in analytical expressions of curvature, in special kinds of symmetries of curves; he introduced polar coordinates, and treated similar and affine curves.

3.2 Transcendental curves. Only in the last chapter of his theory of curves did Euler treat curves that did not depend only upon algebraic functions but also upon so-called transcendental curves: 'Hence a transcendental line, which is what such a curve is called, is defined to be one such that the relationship between the abscissa and the ordinate cannot be expressed by an algebraic equation' (art. 506). He mentioned logarithms and trigonometric functions, but there were innumerable other expressions: 'The number of transcendental curves is much larger than the number of the algebraic curves' (art. 507). But he was not even shocked by curves such $y = (-1)^x$, $y = x^x$ and $x^y = y^x$, as well as the equation $\cos x = x$.

3.3 Surface theory. In quoting Alexis Clairaut Euler referred to curves in space, which were closely related to surface theory. But he did not give a general theory of surfaces; he just treated several points as the surfaces of solids, the intersection of a surface with a plane, especially sections of cylinders, cones and spheres, as well as second-order surfaces and the intersection of two surfaces. This last chapter included the theory of spatial curves. Only in [Euler, 1767] was he to give a definition of the curvature of a surface at a fixed point: if f and g are the radii of curvature of the principal sections, and φ the angle between an arbitrary normal section and a principal section, then the radius r of curvature through that section is given by

$$r = \frac{2fg}{f + g - (f - g)\cos 2\varphi}. \tag{11}$$

4 ON THE IMPACT OF THE *INTRODUCTIO*

All of the mathematics in Euler's book was truly pre-calculus introductory; staple mathematics, whether newly minted by the author here or retrieved and developed from former work by himself and/or others. Thus it was a repository of a mass of useful information about functions, series, curves and surfaces of many kinds. Only one rather specialised topic appeared in the book; the partition of numbers into additive parts (ch. 16), a topic that he had opened up in papers from 1740.

Naturally the book became very well known and used; but such import did not necessarily convey into citations, for the book had became part of the woodwork for training and consultation in analysis and geometry rather than a source for citation. A crest of reception may be determined in France during the 1790s, following the French Revolution and indeed much driven by the educational reforms that had been set in train, especially the new engineering school the *Ecole Polytechnique* (1794). Three years later founder professor of analysis, J.L. Lagrange, published his textbook on analytic functions, a famous work that is the subject of §19; founded upon the algebraically inspired belief that every function could be expanded in a Taylor series, several features of the *Introductio* lay at least in the background of the ensuing discourse. Lagrange was succeeded in 1799 as professor by

S.F. Lacroix, who was then completing a treatment of the calculus and related topics in a massive three-volume *Traité* (1797–1800) that is likewise home to §20. While Lagrange warmed to the algebra of Euler's first volume, Lacroix gave both volumes due attention. Finally, the publication history at the head of this article records a full French translation appearing in 1796–1797; it was prepared by a junior staff member at the school. We also note from there that the book had already appeared fully in German.

Thereafter, as other countries began partially to emulate France in instruction in higher mathematics, further textbooks and treatises began to appear, in various languages, and Euler's *Introductio* steadily became more of the furniture. But references can still be found to it both in textbooks and research work of all kinds, as an author appealed to its authority for some function, series expansion, continued fraction, or property of curve or surface. The book gained the ultimate accolade of being taken for granted.

BIBLIOGRAPHY

Boyer, C. 1956. *History of analytic geometry*, New York: Scripta Mathematica.

Cantor, M. 1898. *Vorlesungen über Geschichte der Mathematik*, vol. 3 (*1668–1758*), Leipzig: Teubner.

Euler, L. *Works*. *Opera omnia*, 3 series, in progress, 1911–; now Basel: Birkhäuser.

Euler, L. 1736. *Mechanica sive motus scientia analyticae expositia*, 2 vols., Saint Petersburg: Academy. [Repr. as *Works*, ser. 2, vols. 1–2.]

Euler, L. 1744. 'De fractionibus continuis dissertatio', *Commentarii Academiae Scientarum Petropolitanae, 9* (1737), 98–137. [Repr. in *Works*, ser. 1, vol. 14, 187–215.]

Euler, L. 1767. 'Recherches sur la courbure des surfaces', *Mémoires de l'Académie des Sciences de Berlin, 16* (1760), 119–143. [Repr. in *Works*, ser. 1, vol. 28, 1–22.]

Euler, L. 1794. 'De formulis differentialibus angularibus maxime irrationalibus quas tamen per logarithmos et arcus circulares integrare licet', in *Institutiones calculi integralis*, vol. 4, Pavia, 183–194. [Repr. in *Works*, ser. 1, vol. 19, 129–140.]

Euler, L. 1983. *Leonhard Euler 1707–1783. Beiträge zu Leben und Werk*, Basel: Birkhäuser.

Fellmann, E.A. 1983. 'Leonhard Euler. Ein Essay über Leben und Werk', in [Euler, 1983], 13–98.

Ferraro, G. 1998. 'Some aspects of Euler's theory of series', *Historia mathematica, 25*, 290–317.

Gelfond, A.O. 1983. 'Über einige charakteristische Züge in den Ideen L. Eulers auf dem Gebiet der mathematischen Analysis und seiner "Einführung in die Analysis des Unendlichen" ', in [Euler, 1983], 99–110.

Kline, M. 1983. 'Euler and infinite series', *Mathematics magazine, 56*, 5, 307–315.

Speiser, A. 1945. 'Vorwort Introductio in analysin infinitorum', in [Euler, *Works*], ser. 1, vol. 9, Leipzig and Berlin: Teubner, vii–xxxii.

Thiele, R. 1982. *Leonhard Euler*, Leipzig: Teubner (*Biographien hervorragender Naturwissenschaftler, Techniker und Mediziner*, vol. 56).

Thiele, R. 2000. 'Frühe Variationsrechnung und Funktionsbegriff', in his (ed.), *Mathesis. Festschrift zum 70. Geburtstag von Matthias Schramm*, Berlin, Diepholz: GNT Verlag, 128–181.

Thiele, R. 2004. 'The mathematics and science of Euler', in *Collection of Kenneth May Lectures delivered before the CSHPM*, New York: Springer, to appear.

Yushkevich, A.P. 1976. 'The concept of function up to the middle of the 19th century', *Archive for history of exact sciences, 16*, 37–85.

CHAPTER 14

LEONHARD EULER, TREATISE ON THE DIFFERENTIAL CALCULUS (1755)

S.S. Demidov

In this book Euler gave a detailed and updated account of the calculus in its Leibnizian tradition. In addition to making many applications to series, functions, the theory of equations and differencing, he modified the theory itself by introducing the differential coefficient.

First publication. Institutiones calculi differentialis cum eius vsu in analysi finitorum ac doctrina serierum, Berlin: Impensis Academiae Imperialis Scientiarum Petropolitanae, 1755. xx + 880 pages.

Later editions. 1) Pavia: Galleati, 1787 [with list of works and obituary]. 2) As Euler, *Opera omnia*, ser. 1, vol. 10 (ed. G. Kowalewski), Basel: Orell Füssli, 1913.

German translation. Vollständige Anleitung zur Differential-Rechnung (trans. J.A.C. Michelsen), 3 vols., Berlin and Libau: Lagarde and Friedrich (vols. 1–2), Lagarde (vol. 3), 1790–1793.

Russian translation. Differentcial'noe ischislenie (trans. and ed. M.Ya. Vygodskii), Moscow and Leningrad: GTTI, 1949.

Partial English translation. Of Part 1 as *Foundations of differential calculus* (trans. J.D. Blanton), New York: Springer, 2000. [Some anachronisms.]

Related articles: Leibniz (§4), MacLaurin (§10), Euler *Introductio* (§13), Lagrange on the calculus (§19), Lacroix (§20).

1 THE SECOND PART OF EULER'S TRILOGY ON MATHEMATICAL ANALYSIS

Euler's 'Differential calculus' (hereafter, 'DC') constitutes the second part of his encyclopedic work on mathematical analysis: the first one was the 'Introduction to the analysis of the Infinitesimals' [Euler, 1748], in two volumes (§13), and the third

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

was the 'Integral calculus' [Euler, 1768–1770], in three volumes. He started to write this book already in Saint Petersburg and finished it around 1750 in Berlin, where it was published under the auspices of the Saint Petersburg Academy of Sciences [Yushkevich and Winter, 1960, 437–438]. The existence of an early Latin manuscript 'Calculi differentialis', conserved in the Archives of the Russian Academy of Sciences in Saint Petersburg (fund 136, inventory 1, opus 183, fols. 1–15) shows that Euler worked over a very long period to present his modern view of the differential calculus. An account of his scientific manuscripts dated this one to the 1730s [Kopelevich et alii, 1962, 41], while A.P. Yushkevich considered that it was written even earlier, around 1727 [Yushkevich, 1983, 161]. We consider that its comparison with the book of 1755 reveals the evolution of the calculus during these 20 years (to a great extent due to Euler himself) and the modification of his orientation: while the manuscript reveals his approach to the infinitesimals as a pupil of Johann Bernoulli, in the book of 1755 he founded the calculus on his own 'calculus of zeros'.

The contents of the book are summarised in Table 1. The exposition, which is very succinct, comprises two Parts, each with its own sequence of numbered chapters and articles. Despite the diversity of the topics and the impressive size, it is a complete, well organized treatise. Many of the results are Euler's own. The first Part is devoted to the differential calculus and its foundations, and the second Part contains applications of the differential calculus related to analysis and algebra. At the end of the first Part and in the last chapters of the second Part he states his intention to write a third Part, devoted to the geometrical applications of the differential calculus; but he never realizes it. The section of its manuscript prepared around 1750 was published only in [Euler, 1862].

2 THE DIFFERENTIAL CALCULUS AND ITS FOUNDATIONS

In the extended introduction Euler explains the purpose of calculus, including, in particular, his famous 'expanded' conception of a mathematical function: 'if some quantities depend on others in such a way as to undergo variation when the latter are varied, then the former are called functions of the latter'. This formulation has an extensive character ('*quae denominatio latissime patet*'); it embraces all the ways by which one quantity can be determined by means of others, and anticipates the definitions of later mathematicians such as N.I. Lobachevsky and J.P.G. Dirichlet. However, in his book Euler's conception is not utilized in practice: functions are mainly considered as analytical expressions, including infinite series. His introduction also includes a very concise and schematic historical essay, a criticism of the foundation of the calculus on the infinitesimals, and a very brief survey of the book's contents.

In its first Part Euler exposes the elements of the calculus of finite differences, indispensable for him to build his version of the calculus. In his view the principal object is not the differential, but the derivative. But unlike Newton this notion is defined by Euler without any use of the concept of velocity but purely arithmetically; as he writes in his introduction, as 'a ratio of vanishing increments, receiving by some functions, when the variable receives a vanishing increment'. For Euler these vanishing in-

Table 1. Summary by Sections of Euler's book. Part 2 starts at Section IIIA.

Sec.	**Chs.**	**Art.**	**P.**	**'Title' or Description**
IA		pp. iv–xx		'Introduction'.
IIA	1–2	1	3	Calculus of finite differences.
IIB	3–4	72	71	Foundation of the differential calculus: calculus of zeros; on the notions of the differential and the integral, on the differentials of higher orders.
IIC	5–8	152	124	Differentiation of functions: algebraic, transcendental, of two variables (on necessary conditions that $P(x, y)\,dx + Q(x, y)\,dy$ be a total differential); successive differentiation of functional expressions (substitution of variables, etc.).
IID	9	281 –327	241	'On differential equations': differentiation of implicit functions, obtaining different differential equations from a given finite equation; expansion of the order of the differential equation to eliminate one of the variables as constant quantities.
IIIA	1	1	281	'On the transformations of series': necessary information from the theory of series, on substitutions; transforming a given series to another one which diverges more quickly.
IIIB	2	19	304	Summation methods of series (differentiation of series, algebraic transformations, etc.).
IIIC	3–4	44	332	On the presentation of finite differences of the function by its derivatives, on Taylor series and some of its applications.
IIID	5–7	103	403	Euler–Maclaurin summation formula and its applications.
IIIE	8	198	515	The development of different functions in series by the method of undetermined coefficients.
IIIF	9	227	546	'On the application of differential calculus to the solution of the equations'; approximate solution of algebraic equation.
IIIG	10–11	250	656	On the application of differential calculus to the study of maximum and minimum of functions.
IIIH	12–13	251	657	On the valuation of the numbers of the real and the imaginary roots of algebraic equations.
IIII	14	337	712	'On the differentials of functions in some particular cases: cases when it is impossible to consider the increment on the function as equivalent to its differential'.
IIIJ	15	355	738	On the indefinite forms $0/0$, ∞/∞, $\infty - \infty$.
IIIK	16–17	367	769	'On the interpolation of the functions determined for the natural values of an argument for the fractional and even irrational values of argument'.
IIIL	18	403 –480	843 –880	'On the application of the differential calculus to the development of fractions'.

crements became zeros and the derivative of the function $f(x)$ is considered as the ratio dy/dx, where $dy = 0$ and $dx = 0$. Nevertheless, Euler does not accept the objection that division by zero has no sense: the ratio $0/0$ could equal a definite number, for 'It is quite clear from the simple arithmetic; everybody knows that a zero multiplied to any number n gives zero, that is to say $n \cdot 0 = 0$, it is why $n : 1 = 0 : 0$' (art. 85). In the differential calculus dy/dx ceases to be indefinite. He presents the rules on the determination of the ratios of the differentials (calculus of zeros) in two forms (arts. 87–97), rejecting the infinitesimals and corresponding to the passage to the limit:

$$a \pm n\,dx = a \text{ and, correspondingly, } a \pm n\,dx/a = 1; \tag{1}$$

$$\text{for } n > m\text{: } a\,dx^m \pm b\,dx^n = a\,dx^m \tag{2}$$

and, correspondingly,

$$\left(a\,dx^m \pm b\,dx^n\right)/a\,dx^m = 1. \tag{3}$$

By considering the finite increment of a function $y(x)$, when its argument x receives a finite increment ω (art. 112), Euler obtains the form

$$\Delta y = P\omega + Q\omega^2 + R\omega^3 + S\omega^4 + \cdots. \tag{4}$$

Using his principle to banish the infinitesimals, he obtains $dy = P\omega$. Now, designating Δy as dy and ω as the constant value dx, he obtains $dy = P\,dx$ (arts. 114, 118). So 'if P is found, the ratio between dx and dy is known' (art. 120) without searching for the limit of $\Delta y/\Delta x$. In such manner Euler obtains in the 5th and 6th chapters of the first Part the derivatives of the power function, the quotient, the logarithmic function, the sine and cosine, and so on. So for the formula

$$d\,\ln x = \ln(x + dx) - \ln x = \ln(1 + dx/x) \tag{5}$$

(for our symbol 'ln' he used 'l') he utilizes the expansion

$$\ln(1 + z) = z - z^2/2 + z^3/3 - \cdots, \tag{6}$$

obtained in his earlier 'Introduction' [1748, vol. 1, art. 123], and, replacing z by dx/x, he obtains $d\ln x = dx/x$.

In a notable passage Euler showed that the indeterminacy of higher-order differentials could be eliminated by assigning a constant value to dx, so that $d\,dx = dd\,dx = \cdots = 0$. The variable x was now given a special status, equivalent to our practice of assigning as independent, with the others functionally dependent upon it: in particular,

$$dy = p\,dx; \quad \text{thus } d\,dy = dp\,dx + p\,d\,dx = q(dx)^2, \text{ where } dp = q\,dx, \tag{7}$$

and so on for higher orders $dd\,dy, \ldots$ (arts. 128–130, 251–261). By introducing the functions p, q, which are to be called 'differential coefficients' by S.F. Lacroix (§20), he anticipated J.L. Lagrange's later emphasis on the central role of the derivative (§19.2) [Bos, 1974, pt. 5; Ferraro, 2004].

Euler exposes in details the methods for the differentiation of functions of one and many variables, the rules of the differentiation of functions of some quantities depending on one argument, proves the theorem on the independence of the value of partial derivatives from the order of differentiation, demonstrates the theorem on homogeneous functions. He also presents the necessary condition that the expression $P(x, y)\,dx + Q(x, y)\,dy$ is a total differential, namely

$$\left(\frac{dP}{dy}\right) = \left(\frac{dQ}{dx}\right). \tag{8}$$

3 APPLICATIONS OF THE DIFFERENTIAL CALCULUS

In the second Part, Euler examines the applications of the differential calculus in some problems related to analyses and to algebra. He derives Taylor series and gives several applications, in particular to the development of various functions in power series, the determination of its numerical values, and the arithmetical solution of algebraic equations (by a method of determining more exact values of its solution after an initial weak approximation). He deduces the Euler–Maclaurin summation formula (which he had found earlier: Mills [1985]), and relates the partial sums to the integrals and the derivatives of its general term having the form

$$S(x) = \int z\,dx + \frac{1}{2}z + \frac{B_2}{1\cdot 2}\frac{dz}{dx} + \frac{B_4}{1\cdot 2\cdot 3\cdot 4}\frac{d^3z}{dx^3} + \frac{B_6}{1\cdot 2\cdot 3\cdot 4\cdot 5\cdot 6}\frac{d^5z}{dx^5} + \cdots, \tag{9}$$

where B_2, B_4, $B_6, \ldots$ are the Bernoulli numbers (Euler uses another symbol for them), $S(x)$ is the sum of the x first terms of the series, $z(x)$ is its general term, and $S(0) = z(0) = 0$. Euler utilizes his formula for various numerical series and receives many wonderful and elegant results (chs. 5–7). The use of asymptotic series presents a special interest; from them he obtains the value of the constant included in this formula. In particular, for the harmonic series this constant became 'Euler constant'.

Another example is this presentation (art. 156) of π:

$$\pi = \sum_{k=1}^{n} \frac{4n}{n^2 + k^2} + \frac{1}{4n} + \frac{B_2}{2\cdot 2\cdot n^2} - \frac{B_6}{2^3\cdot 6\cdot n^{10}} + \frac{B_{10}}{2^5\cdot 10\cdot n^{10}} - \cdots. \tag{10}$$

Realising the divergence of this series, in his calculation of π Euler restricts the series to $n = 5$, terminating the summation in time. As one result he finds π to twelve digits.

Euler paid great attention to the extrema of a function $y = f(x)$. He deduces the necessary and sufficient conditions to find them: thus, he studies the sign of the difference

$$f(x \pm \alpha) - f(x) = \pm\alpha\frac{dy}{dx} + \frac{\alpha^2}{2}\frac{d^2y}{dx^2} \pm \cdots \tag{11}$$

in a sufficiently small neighbourhood of the corresponding value of the argument. At the same time he utilizes a proposition (which is rigorous only under some restrictions) anal-

ogous to the principle of banishing the infinitesimals: for sufficiently small α the absolute value of some term of the series is bigger of the sum of all the rest terms (art. 254). This proposition became one of the principles in Lagrange's foundations of analysis (§19). The use for the value of the forward difference $(f(x \pm \alpha) - f(x))$ of the Taylor series with remainder instead of the infinite Taylor series allowed a more precise use of the series [Yushkevich, 1968, 149].

Euler also studies the extremities of multiple-valued functions and for functions of many variables. Astonishing is a major mistake, when he considers that a function of two variables must have its maximum (or its minimum) if it have the maximum (or minimum) accordingly any argument when the other one is constant (art. 290); Lagrange was to correct this. Among other applications of the calculus considered by Euler are questions concerning the number of real and imaginary roots of an algebraic equation (his results were based on the assumption that between two roots of the equation $f(x) = 0$ there must exist a root of the equation $f'(x) = 0$), and also the analysis of the indefinite forms $0/0$, ∞/∞ and $\infty - \infty$ (ch. 15).

4 GENERAL REMARKS

We stress an important aspect of the mathematical mentality of the 18th century which clearly appeared in Euler's works, especially in this book. For the mathematicians of his epoch the mathematical notions and operations originated from experience, and through it they were determined. Hence it was certain that the establishment of a mathematical statement could be made not only from another statements considered as true, but also, for example, from some physical considerations. So, an argument in the famous discussion from the mid 1740s onwards on the nature of arbitrary functions in the solution of the equation of the vibrating string, where Euler was a principal protagonist, could have not only mathematical, but also physical nature. This solution is not (only) the concept that can be determined in such or such manner, but is an expression of the objective nature of the physical phenomena itself. In this discussion Euler's position was based on physical considerations [Truesdell, 1960, pt. 3].

This view is the origin of Euler's attitude to incomplete inductions, unexpected for a modern reader and more appropriate to a physicist. Without any restriction, Euler uses it even in his DC. For him a regularity determining $n!$ for entire numbers is quite sufficient to find $(1/2)! = \sqrt{\pi}/2$ (pt. 2, art. 402). M.Ya. Vygodskii thought that 'for him "an incomplete induction" is sufficiently convincing' [1949, 23]. But Yushkevich did not agree with this interpretation; for him Euler used an incomplete induction rather as 'an instrument of scientific research' [1968, 113]. Certainly Euler gave the 'complete' induction a higher status than the incomplete; in number theory the incomplete induction is not sufficient (head of [Euler, 1741]). In such a branch as mathematical analysis, which was not rigorously defined and entirely free from mechanics and geometry (which gave birth to it), it was possible, according to Euler, to use 'incomplete induction' in the proof: he gives many of such examples.

At the same time, the foundations of analysis were made manifest as an independent mathematical discipline in Euler's works, in particular his DC. The lack of geometrical

and mechanical applications but even geometrical and mechanical illustrations constitutes a very important special feature of the book. It contains, in its introduction, only one physical example, about the trajectory of a cannon ball! The entire exposition has an abstract arithmetico-algebraic character, and not without pride Euler wrote in its introduction: 'all the exposition is bounded in the frame of pure analysis, so for the exposition of all the rules we did not use even one figure'. It is possible to discuss the methodical quality of such a kind of exposition, but it was very important for the future development of mathematics. In a nearest perspective such an exposition, thanks to the emancipation of the analysis from geometrical and mechanical ideas, liberates it from erroneous conclusions imposed by them. In a distant perspective such an exposition, which expressed the objective tendency to establish the analysis as an independent discipline, was a model for the future books on analysis and prepared the way for its arithmetization. At the same time DC became a mine of concepts for several generations of mathematicians in the 18th and 19th centuries; for example, asymptotic developments, divergent series, and the zeta-function.

BIBLIOGRAPHY

Bos, H.J.M. 1974. 'Differentials, higher-order differentials and the derivative in the Leibnizian calculus', *Archive for history of exact sciences, 14,* 1–90.

Euler, L. 1741. 'Theorematum quorundum ad numeros primos', *Commentarii Academiae Scientiarum Petropolitanae, 8* (1736), 141–146. [Repr. in *Opera omnia*, ser. 1, vol. 2, 33–37.]

Euler, L. 1748. *Introductio in analysin infinitorum,* 2 vols., Lausanne: Bousquet. [Repr. as *Opera omnia*, ser. 1, vols. 8–9. See §13.]

Euler, L. 1768–1770. *Institutiones calculi integralis,* 3 vols., Saint Petersburg: Academy. [Repr. as *Opera omnia*, ser. 1, vols. 11–13.]

Euler, L. 1862. *Opera posthuma,* vol. 1, Saint Petersburg: Academy.

Euler, L. 1983. *Leonhard Euler. Beiträge zu Leben und Werk,* Basel: Birkhäuser.

Ferraro, G. 2004. 'Differentials and differential coefficients in the Eulerian foundations of the calculus', *Historia mathematica, 31,* 34–61.

Ferraro, G. and Panza, M. 2003. 'Developing into series and returning from series', *Historia mathematica, 30,* 17–46,

Kopelevich, Yu.Kh., Krutikova, M.V., Mikhailov, G.K. and Raskin, N.M. 1962. *Rukopisnye materialy Leonarda Eilera v Arkhive Akademii Nauk SSSR. Nauchnoe opisanie* ['Manuscripts of L. Euler in the Archives of the Academy of Sciences of USSR. Scientific description'], vol. 1, Moscow and Leningrad: Izdatel'stvo AN SSSR.

Mills, S. 1985. 'The independent derivations by Leonhard Euler and Colin MacLaurin of the Euler–MacLaurin summation formula', *Archive for history of exact sciences, 33,* 1–14.

Truesdell, C.A., III 1960. *The rational mechanics of flexible or elastic bodies 1638–1788,* Zurich: Orell Füssli [as Euler *Opera omnia,* ser. 2, vol. 11, pt. 2].

Vivanti, G. 1908. 'Infinitesimalrechnung', in M. Cantor (ed.), *Vorlesungen über Geschichte der Mathematik,* vol. 4, Leipzig: Teubner, 639–689.

Vygodskii, M.Ya. 1949. 'Vstupitel'noe slovo k "Differentcial'nomu ischisleniyu" L. Eilera' [Introduction to L. Euler's 'Differential Calculus'], in Euler, *Differentcial'noe ischislenie,* Moscow and Leningrad: GTTI.

Yushkevich, A.P. and Winter, E. 1960. 'O perepiske Leonarda Eilera s Peterburgskoi Akademiei Nauk v 1741–1757' ['On the correspondence of Leonhard Euler with the Petersburg Academy of Sciences in 1741–1757'], in *Trudy Instituta Istorii Estestvoznaniya i Tekhniki, 34,* 428–491.

Yushkevich, A.P. 1968. *Istoriya matematiki v Rossii do 1917 goda* [‘History of mathematics in Russia up to 1917’], Moscow: Nauka.

Yushkevich, A.P. 1983. ‘L. Euler’s unpublished manuscript “Calculus differentialis” ’, in [Euler, 1983], 161–170.

CHAPTER 15

THOMAS BAYES, *AN ESSAY TOWARDS SOLVING A PROBLEM IN THE DOCTRINE OF CHANCES* (1764)

A.I. Dale

In this paper Bayes published his theorem on prior and posterior probabilities. While its reception was slow, it has led to the widespread use of 'Bayesian' to describe an influential construal of types of probability and statistics statements.

First publication. *Philosophical transactions of the Royal Society of London, 53* (1763: publ. 1764), 370–418.

Reprints. 1) As *A method of calculating the exact probability of all conclusions founded on induction*. London: 1764. 2) In *Facsimiles of two papers by Bayes* (ed. E.C. Molina and W.E. Deming), Washington: The Graduate School, The Department of Agriculture, 1940 [repr. New York: Hafner, 1963]. 3) Corrected version in *Biometrika*, *45* (1958), 296–315, following [Barnard, 1958]; this version repr. as appendix to S.J. Press, *Bayesian statistics: principles, models, and applications*, New York: Wiley, 1989. 4) In ed. R. Swinburne, *Bayes's theorem*, Oxford: Oxford University Press, 2002, 117–149.

German translation. *Versuch zur Lösung eines Problems der Wahrscheinlichkeit von Thomas Bayes* (trans. and ed. H.E. Timerding), Leipzig: Engelmann, 1908 (*Ostwald's Klassiker der exakten Wissenschaften*, no. 169).

French translation. *Essai en vue de résoudre un problème de la doctrine des chances* (trans. J.P. Clero, preface by B. Bru), *Cahiers d'Histoire et de Philosophie des Sciences*, no. 18 (1988), Paris: Société Française d'Histoire des Sciences et des Techniques. [Includes a photocopy of the original.]

Related articles: Jakob Bernoulli (§6), De Moivre (§7), Laplace on probability (§24).

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

1 BIOGRAPHY

The scion of a respectable line of cutlers in Sheffield, England, Thomas Bayes, the eldest of seven children, was born in Bovingdon, Hertfordshire, in 1702. Both place and date of birth must, however, be viewed with some caution: his birthdate is found by subtracting his age at death from the year of death (both recorded on the Bayes family vault), while the birthplace is derived from the known presence of his parents in that town in the late 17th century and their later residence in London. His father Joshua, having been ordained in London in 1694, moved with his wife Anne (née Carpenter) to Box Lane, Bovingdon. In 1707 the family returned to London, where, after some years, Joshua became minister at the Presbyterian Chapel in Leather Lane, remaining there until his death in 1746.

After studying theology at Edinburgh University, Bayes spent some time in London before accepting a position as minister at the Mount Sion Meeting-house in Tunbridge Wells, Kent, where he was to remain until his death on 7 April 1761. His remains were taken to London, and interred in the family vault in Bunhill Fields burial-ground in Moorgate. This vault, which has been repaired a number of times over the years, now carries the information that it was restored 'In recognition of Thomas Bayes's important work in probability [...] in 1969 with contributions received from statisticians throughout the world'. On his life and work, see [Dale, 2003].

Bayes's first published work was a tract entitled *Divine Benevolence, or, an attempt to prove that the principal end of the divine providence and government is the happiness of his creatures* (1731). This was followed by *An Introduction to the doctrine of fluxions* (1736), a rebuttal to George Berkeley's *The Analyst: or, a discourse addressed to an infidel mathematician* (§8). Both works were published anonymously. Although Bayes published nothing else, manuscripts in the Stanhope of Chevening papers, recently discovered by David Bellhouse in the county archives in Maidstone, Kent, indicate that Bayes acted as an 'adviser' on mathematical writings: other manuscripts are to be found in the Royal Society and in the Equitable Life Assurance Society in London [Bellhouse, 2001].

Among the Royal Society manuscripts is a letter from Bayes to John Canton commenting on Thomas Simpson's suggestion, as expressed in his paper in the *Philosophical Transactions* in 1755, that the error in astronomical observations could be reduced by taking several measurements rather than just one. Bayes's objection was that this would not do if the measuring instrument were inaccurate—and he was also unhappy with the recommendation that errors in excess as in defect should be taken as equiprobable.

After Bayes's death his friend Richard Price (1723–1791) arranged for the forwarding, for reading and publication, of a number of papers to the Royal Society; Bayes had been elected Fellow in 1742, his proposers being men of considerable scientific weight such as Martin Folkes, John Eames and Philip, Earl Stanhope. Indeed, D.O. Thomas suggests that it was his publication of Bayes's papers that set in train Price's 'increasing involvement in insurance, demography, and financial and political reform' [Thomas, 1977, 128].

The first of these papers was devoted to the divergence of the Stirling–de Moivre series expansion of $\log z!$ as

$$(1/2)\log 2\pi + \big(z + (1/2)\big)\log z - \big[z - (1/12z) + \big(1/360z^3\big) - \big(1/1260z^5\big) + \&\text{c.}\big]. \quad (1)$$

Although Leonhard Euler had shown, some six years before Bayes's death, that the series failed to converge for $z = 1$, Bayes was apparently the first to note the general lack of convergence.

Bayes's two published tracts show him as a man to be taken seriously by both his theological and his mathematical contemporaries. *Divine Benevolence* prompted a tract in rebuttal from Henry Grove, and Philip Doddridge, the head of a well-known Nonconformist academy, gave lectures on it. The tract on fluxions would certainly have brought Bayes to Earl Stanhope's notice, and it was perhaps instrumental in the latter's using Bayes as a mathematical referee. However it is for posthumously published work on probability that Bayes is remembered, and it is to this that we now turn.

2 BAYES'S WORK ON CHANCES

Bayes's main work, the second of the posthumous papers mentioned above, is the important 'An essay towards solving a problem in the doctrine of chances'. Here are to be found the origins of the modern ideas of prior and posterior probabilities, concepts on which the whole theory of Bayesian statistics is based.

The problem with which Bayes concerned himself was the following. Given the number of times in which an unknown event has happened and failed: required the chance that the probability of its happening in a single trial lies somewhere between any two degrees of probability that can be named. The solution, expressed in modern notation, is given in the tenth proposition:

Let x be the (prior) probability of an unknown event A. Then

$$\Pr[x_1 < x < x_2 \mid A \text{ has happened } p \text{ times and failed } q \text{ times in } p+q \text{ trials}]$$
$$= \int_{x_1}^{x_2} \binom{p+q}{p} x^p (1-x)^q \, dx \Big/ \int_0^1 \binom{p+q}{p} x^p (1-x)^q \, dx. \tag{2}$$

Bayes's solution of his problem was given essentially as a ratio of areas and evaluated as an infinite series, and not in terms of the incomplete beta integral given above.

The result most commonly known today as Bayes's Theorem is however not that given above, but rather

$$\Pr[B_i \mid A] = \Pr[A \mid B_i] \Pr[B_i] \Big/ \sum_i \Pr[A \mid B_i] \Pr[B_i]. \tag{3}$$

This version does not appear in Bayes's work itself, but is found for the first time in a paper by Pierre-Simon Laplace (1749–1827), 'Mémoire sur la probabilité des causes par les événements', where the notion of inverse probability was presented (Laplace [1774]: compare §24.2). Bayes's geometric approach yields a result for continuous probabilities, while Laplace begins with an urn containing a finite number of balls and *then* passes to the case of an urn with an infinite number of balls. It would appear that the *Mémoire* was written in ignorance of the *Essay*, for it was only in later papers that Laplace acknowledged Bayes's seminal contribution [Gillies, 1987]. We would thus still have 'Bayes's Theorem' today even if Price had not submitted Bayes's manuscript to the Royal Society.

One of the main philosophical, as opposed to the purely mathematical, aspects of the *Essay*, is the question of *time*. In his introduction to the *Essay* Price notes that Bayes gave a specific definition of *chance* (or *probability*), 'which in common language is used in different senses by persons of different opinions, and according as it is applied to *past* or *future* facts'. Bayes's definition gives the probability of any event as 'the ratio between the value at which an expectation depending on the happening of the event ought to be computed, and the value of the thing expected upon its happening'. The matter is of particular importance in the interpretation of conditional probability as it arises in the third, fourth and fifth propositions in the *Essay*. We shall say more on this matter later: for a detailed discussion of the different treatments needed depending upon whether one is concerned with the order in which two events have occurred (or will occur) or the order in which one learns they have occurred, see the Clero–Bru translation of the *Essay* listed above.

Basic to Bayes's proof of his main result is the following postulate: suppose that a level square table be made in such a way that a ball W thrown upon it has the same probability of coming to rest at any particular point as at any other point. Suppose further that, after this first ball has been thrown onto the table, a further $(p+q)$ throws are made with a second ball, and that each of these throws results in the success or failure of an event M according as the second ball is nearer to or further from a specified side of the table than the first ball. That is, in modern terminology, a) a single value x is drawn from a uniform distribution concentrated on $[0, 1]$, and b) a sequence of Bernoulli trials, with success probability x, is generated.

Bayes's assumption of a uniform distribution as a prior when one is in a state of ignorance is not one that has enjoyed universal acceptance, and many authors have suggested alternative distributions for the representation of such ignorance (see, for example, [Jeffreys, 1961]). In some cases the choice of a prior is less important than the data in the determination of the posterior distribution: for instance, L.J. Savage writes in the context of precise measurement: 'This is the kind of measurement we have when the data are so incisive as to overwhelm the initial opinion, thus bringing a great variety of realistic initial opinions to practically the same conclusion' [Savage, 1962, 29]. On the other hand, A. O'Hagan has noted that 'the prior distribution can be made strong enough to overwhelm any data' [1994, art. 3.27].

Three rules were given in the *Essay* for the obtaining of bounds to the exact probability required, their proofs being held over to the sequel paper on 'A demonstration of the second rule in the essay towards the solution of a problem in the doctrine of chances, with improvements' obtained by Price of Bayes's bounds [Bayes and Price, 1765]. The first rule may be written as follows:

$$\Pr[x_1 < x < x_2 \mid p, q] = (n+1)\binom{n}{p}\sum_{i=0}^{q}\binom{q}{i}\left(\frac{1}{p+1+i}\right)\left(x_2^{p+1+i} - x_1^{p+1+i}\right). \quad (4)$$

This form of the rule was to be used for large p and small q, something similar holding for q large and p small. The second and third rules were to be used when p and q were both large.

To some extent, the Rules and their proofs are responsible for the difficulty of the *Essay* and the *Demonstration*. Both Bayes and Price, after giving various approximations to the

incomplete beta integral, examined the accuracy of their approximations, especial attention being paid to the maximum error that may be incurred in the making of such approximations [Hald, 1990, 140].

Abraham de Moivre had proved, in 1733, that the *symmetric* binomial distribution, with probability density $\binom{n}{x}(1/2)^n$, tends, as $n \to \infty$, to the normal distribution, and that the latter could therefore be used as an approximation to a cumulative binomial (§7.6). He later also showed that the same limit obtained for the *skew* binomial.

In the course of this work de Moivre essentially found the expansion

$$\frac{2}{\sqrt{\pi}} \int_0^t \exp\bigl(-u^2\bigr)\, du = \frac{2}{\sqrt{\pi}} \sum_{k=0}^{\infty} \frac{(-1)^k t^{2k+1}}{k!(2k+1)}, \tag{5}$$

while Price derived the same series in the *Demonstration* as an approximation to the posterior distribution, the latter being that arrived at in the *Essay*. This posterior is a beta distribution, and Price's results thus anticipated those published by Laplace in a 'Mémoire sur les probabilités' in 1781.

In the results given in both the *Essay* and the *Demonstration* relations between integrals were verified by showing that the corresponding relations held between the integrands. These latter relations in turn were shown to obtain by examination of the derivatives of the integrands and the use of monotonicity, a technique that had been profitably employed by Bayes in *An Introduction to the doctrine of fluxions.*

3 PRICE'S APPENDIX

An Appendix to the *Essay* was provided by Richard Price, in which a prospective use of the results of the *Essay* is made. Almost from the start of the Appendix Price applies the results of the *Essay* to the occurrence of future events, writing:

> Let us first suppose, of such an event as that called M in the essay, or an event about the probability of which, antecedently to trials, we know nothing, that it has happened *once*, and that it is enquired what conclusion we may draw from hence with respect to the probability of it's happening on a *second* trial. The answer is that there would be an odds of three to one for somewhat more than an even chance that it would happen on a second trial.

Direct application of Bayes's first rule yields the desired solution, and Price then writes:

> [...] which shews the chance there is that the probability of an event that has happened once lies somewhere between 1 and (1/2); or (which is the same) the odds that it is somewhat more than an even chance that it will happen on a second trial.

While a solution to Price's problem might be obtained by using Laplace's Rule of Succession (e.g. if an event has occurred m times in n trials, the probability that it will occur on the next trial is $(m+1)/(n+2)$) in terms of which the probability of a second occurrence

of the event M would be

$$\int_0^1 x^2\,dx \Big/ \int_0^1 x\,dx = \frac{2}{3}, \tag{6}$$

I do not think that this would be a correct interpretation of the question. For no cognisance would be taken of the requirement that there should be 'more than an even chance that it will happen on a second trial'. It is, however, possible to obtain (6) from Bayes's theory by an appropriate interpretation [Dale, 1999, sect. 4.6].

In another example, one that is noteworthy in view of the important role played therein by the initial event, Price supposes there to be a die of unknown number of faces and unknown constitution (we may suppose the faces to be numbered $n_1, n_2, \ldots, n_k$, not necessarily distinct). If the face n_i (say) appears on the first throw of the die, then we know only that the die has this face. It is only now, i.e. *after* the first throw, that we find ourselves being able to use the results of the *Essay*, and the occurrence of n_i in any future trial is then an event of whose probability we are completely ignorant. If the face n_i appears again on the second trial, then by a previous example in the Appendix, the odds will be three to one on that n_i is favoured—either through being more numerous, or (equivalently) because of the way the die is constituted.

Price now turns to the problem of the probability of the rising of the Sun. Proceeding as in the die-tossing example, he notes that the first sinking of the sun a sentient person who has newly arrived in this world would see would leave him 'entirely ignorant whether he should ever see it again'. Thus, according to Price, 'let him see a second appearance or one *return* of the Sun, and an expectation would be raised in him of a second return, and he might know that there was an odds of 3 to 1 for *some* probability of this', and this is then extended to the case of several occurrences.

After remarking on the probability of causes, Price notes that 'The foregoing calculations further shew us the uses and defects of the rules laid down in the essay'. The defects, as noted by Price, seem to be that the second and third rules 'do not give us the required chances within such narrow limits as could be wished'. These limits, however, contract as q increases with respect to p, the exact solution being given by the second rule when $p = q$.

While Price is perhaps correctly applying Bayes's result in the Appendix, he is applying it to future events. There is no explicit mention in the *Essay* of the applicability of the result to the case of a 'single throw' *after* experience, and it has been suggested elsewhere [Dale, 1999, sect. 4.6] that Bayes's result is in accord with *not* interpreting this 'single trial' in a prospective sense. However it is not obvious from the *Essay* itself that Bayes meant his result to be used only in a retrospective context: Price in fact writes quite explicitly in his introduction that Bayes's original intent was to find the probability of an event given a number of occurrences and failures.

One might also note that Price passes easily from the application of probability in games of chance (his die-tossing example) to its use in connexion with physical phenomena (the problem of the rising of the Sun). The matters raised by Price (both in the Appendix and in his introductory letter) are in fact qualitatively different to those considered in the *Essay*. Whether the notion that is applicable in the case of the tossing of a die is also applicable in the case of natural phenomena could be debated: the analogy would be rejected by some,

while others would perhaps accept it in connexion with matters such as birth ratios but not accept it—or at least query its fitness—in matters such as the lottery example discussed by Price here.

In addition to the remarkable probabilistic result examined here, a number of lesser gems may be glimpsed. Thus we have a clear discussion of the binomial distribution, and probing even further one finds, as [Hailperin, 1996, 14] notes, '(implicitly) the first occurrence of a probability logic result involving conditional probability'. While the mathematician should also be interested in the evaluation of the incomplete beta-function, he will note too the use of approximations to various integrals given in the *Essay* and the *Demonstration* by both Bayes and Price, and the attention paid to the investigation into the error incurred in the making of such approximations.

4 A POSTHUMOUS PUBLICATION

Various suggestions have been proposed for the *Essay* not having been published by the author. The most often assigned cause is modesty, something that seems first to have been attributed to Bayes by Price's nephew, William Morgan who, in his biography of his uncle wrote [Morgan, 1815, 24]:

> On the death of his friend Mr. Bayes of Tunbridge Wells in the year 1761 he [that is, Price] was requested by the relatives of that truly ingenious man, to examine the papers which he had written on different subjects, and which his own modesty would never suffer him to make public.

More recently [Good, 1988] has suggested three possible reasons for non-publication: (a) the tacit assumption of a discrete uniform prior for the number of successes implies that the (physical) probability of a success in each trial has a continuous uniform prior, (b) these two priors are essentially equivalent when the number of trials is large and (c) the first ball is essentially a red herring. Stigler finds a possible reason in the difficulty of evaluating the integral in the eighth proposition [1986, 130], while more recently he suggests that Bayes deferred publication because of 'the lack of an accepted standard of reference that could tell [him] how close to certainty is "good enough" ' [1999, 375].

5 IMPACT AND INFLUENCE OF THE WORK

In the introduction to the first volume of their *Breakthroughs in statistics* [Kotz and Johnson, 1992] listed eleven works, up to and including Francis Galton's *Natural inheritance*, that have had lasting and fundamental effects on the direction of statistical thought and practice. One of these is Bayes's *Essay*.

Despite the use of inverse probability to be found in the work of Laplace and other 19th-century writers, interest in Bayes's work, and realization of its importance, did not really manifest itself until the work of authors like I.J. Good, H. Jeffreys, D.V. Lindley, F.P. Ramsey and L.J. Savage in the middle third of the 20th century (compare §67.3). As a result of this work, Bayesian statistics rapidly became a serious contender to 'classical' or sampling-theory statistics. The Bayesian approach has a number of features that distinguish

it from the more classical rubric: the incorporation of prior information, the fact that all probabilities are subjective, the self-consistency of the method, and the avoidance of having to invent statistical methods.

The fundamental problem in statistics is inference. The Bayesian approach to this matter considers prior beliefs about possible hypotheses (prior to any experimentation, that is) and modifies these in the light of relevant data, yielding posterior beliefs. More specifically, inference about a parameter θ is effected by using Bayes's result to find a posterior density $f(\theta|x)$ from a prior density $f(\theta)$ and a likelihood $f(x|\theta)$. This then allows the answering of questions like 'Having obtained data x, what can be said about parameter θ?'. Note that Bayes himself was concerned with inference only about a 'degree of probability', and not about an arbitrary parameter.

In a paper read at the sesquicentennial meetings of the Royal Statistical Society, [Newell, 1984] pointed out that what Florence Nightingale achieved in hospital design effectively prevented any further development in this field for several decades. In the light of this assertion the statistician might almost be pleased that Bayes's *Essay* received scant attention until the 20th century.

BIBLIOGRAPHY

Bayes, T. and Price, R. 1765. 'A demonstration of the second rule in the essay towards the solution of a problem in the doctrine of chances, with improvements', *Philosophical transactions of the Royal Society of London, 54*, 296–325.

Barnard, G.A. 1958. 'Thomas Bayes—a biographical note', *Biometrika, 45*, 293–295.

Bellhouse, D.R. 2001. 'On some recently discovered manuscripts of Thomas Bayes', *Historia mathematica, 29*, 383–394.

Dale, A.I. 1999. *A history of inverse probability from Thomas Bayes to Karl Pearson*, 2nd ed., New York: Springer.

Dale, A.I. 2003. *Most honourable remembrance. The life and work of Thomas Bayes*, New York: Springer.

Gillies, D.A. 1987. 'Was Bayes a Bayesian?', *Historia mathematica, 14*, 325–346.

Good, I.J. 1988. 'Bayes's red billiard ball is also a herring, and why Bayes withheld publication', *Journal of statistical computing and simulation, 29*, 335–340.

Hailperin, T. 1996. *Sentential probability logic. Origins, development, current status, and technical applications*. London: Associated University Presses.

Hald, A. 1990. 'Evaluations of the beta probability integral by Bayes and Laplace', *Archive for history of exact sciences, 41*, 139–156.

Jeffreys, H. 1961. *Theory of probability*, 3rd ed., Oxford: Clarendon Press.

Kotz, S. and Johnson, N.L. (eds.) 1992. *Breakthroughs in statistics*, 2 vols., New York: Springer.

Laplace, P.S. 1774. 'Mémoire sur la probabilité des causes par les événements', *Mémoires de l'académie royale des sciences de Paris (savants étrangers) 6*, 621–656.

Lindley, D.V. 1972. *Bayesian statistics, a review*, Philadelphia: Society for Industrial and Applied Mathematics.

Morgan, W. 1815. *Memoirs of the Life of The Rev. Richard Price, D.D. F.R.S.*, London: R. Hunter.

Newell, D.J. 1984. 'Present position and potential developments: some personal views. Medical statistics', *Journal of the Royal Statistical Society, A147*, 186–197.

O'Hagan, A. 1994. *Kendall's advanced theory of statistics*, vol. 2B, *Bayesian inference*, Cambridge: Cambridge University Press, for Edward Arnold.

Savage, L.J. and others. 1962. *The foundations of statistical inference. A discussion*, 2nd impression, London: Methuen.
Stigler, S.M. 1986. *The history of statistics. The measurement of uncertainty before 1900*, Cambridge, MA, and London: The Belknap Press of Harvard University Press.
Stigler, S.M. 1999. *Statistics on the table: the history of statistical concepts and methods*, Cambridge, MA: Harvard University Press.
Thomas, D.O. 1977. *The honest mind. The thought and work of Richard Price*, Oxford: Clarendon Press.

CHAPTER 16

JOSEPH LOUIS LAGRANGE, *MÉCHANIQUE ANALITIQUE*, FIRST EDITION (1788)

Helmut Pulte

This is the first textbook to treat theoretical mechanics in a purely analytic way. Its mathematical importance stems mainly from the application of Lagrange's new formalization of the calculus of variations (the δ-calculus), and its significance for rational mechanics from the fact that it summarizes, for the first time in a logically coherent way, the conformity of the Newtonian and continental mechanics of the 18th century on the basis of general variational principles.

First publication. Paris: Desaint, 1788. xii + 512 pages.

Second edition. Mécanique analytique, 2 vols., Paris: Courcier, 1811–1815.

Third edition. 2 vols. (ed. J.L.F. Bertrand), Paris: Mallet-Bachelier, 1853–1855. [Basically the second edition.]

Fourth edition. As *Oeuvres*, vols. 11–12 (ed. J.-A. Serret and G. Darboux), Paris: Gauthier–Villars, 1888–1889. [Photorepr. in one volume: Hildesheim and New York: Olms, 1973.]

English translation of the 2nd edition. *Analytical mechanics* (trans. A. Boissonnade and V.N. Vagliente), Dordrecht: Kluwer, 1997.

German translations. Of the 1st ed.: *Analytische Mechanik* (trans. F.W.A. Murhard), Göttingen: Vandenhoeck und Ruprecht, 1797. Of the 4th ed.: *Analytische Mechanik* (trans. H. Servus), Berlin: Springer, 1887.

Russian translation of the 2nd ed. Analiticheskaya mekhanika (trans. V.S. Hochmana), Leningrad: ONTI, 1938. [2nd ed. Moscow and Leningrad: GTTI, 1950.]

Related articles: Newton (§5), D'Alembert (§11), Euler on curves (§12), Lagrange on functions (§19), Laplace on astronomy (§18), Hertz (§52).

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

1 OUTLINE OF LAGRANGE'S SCIENTIFIC BIOGRAPHY

Joseph Louis Lagrange (1736–1813) was born at Turin (Piedmont) in Italy and baptized as Guiseppe Lodovico Lagrangia. He himself wrote his surname de la Grange, La Grange, or La Grange Tournier in acknowledgment of the French origin of his family [Sarton, 1944; Itard, 1973]. The year of his birth saw the publication of Leonard Euler's (1707–1783) first mathematical textbook, *Mechanica sive motus scientia analytice exposita* [Euler, 1736], a work that can be perceived as a 'semi-analytic forerunner' of the *Méchanique analitique*.

Lagrange's education and early career took place entirely in his home town of Turin, where he spent the first three decades of his life [Borgato and Pepe, 1987]. As early as 1754, at the age of 18, he corresponded with Euler and Giullio di Fagnano (1682–1766) on mathematical questions. He took up a teaching appointment at the Royal College of Gunnery in Turin in 1755 and in the following year became a Foreign Member of the Berlin Academy. Even at this early stage he was developing his δ-calculus, which opened up a new approach to the calculus of variations without recourse to Euler's geometric considerations. Soon afterwards, Lagrange co-founded a scientific society, which later grew into the Royal Academy of Science of Turin. Its journal, *Miscellanea Taurinensia*, first appeared in 1759 and contained some of his most important work on the calculus of variations and analytical mechanics [Lagrange, 1759, 1760a, 1760b, 1770; see Fraser, 1983].

An important influence on his subsequent career was his friendship with Jean le Rond D'Alembert (1717–1783), whom he met during a visit to Paris 1763 and with whom he kept in close touch until his death. It was on D'Alembert's recommendation that Friedrich II (1712–1786) appointed Lagrange director of mathematics classes of the Berlin Academy, a post he relinquished on the death of Friedrich in 1787.

Lagrange produced the *Méchanique analitique* during his time in Berlin. He referred as early as 1756 and 1759 to an almost complete textbook of mechanics, now lost; a later draft first saw the light of day in 1764 [Lagrange, 1764]. But it was not until the end of 1782 that Lagrange seems to have put the textbook into an essentially complete form, and the publication of the book was delayed a further six years [Pulte, 1989, 231].

In 1787, a year before the publication of the *Méchanique analitique*, Lagrange had taken up an appointment as *Pensionnaire vétéran* at the *Académie des Sciénces* in Paris at the invitation of Louis XVI. Following the revolution, which saw the closure of the old Royal Academy, in 1794–1975 he became one of the founder professors of the (short-lived) *Ecole Normale* and of the *Ecole Polytechnique*. It is thus no accident that his most important and influential didactic work in mathematics, such as the *Théorie des fonctions analytiques* ([Lagrange, 1797, 1813]; see §19), was done at this time [Grattan-Guinness, 1990a, vol. 1, 107–109]. Although Lagrange suffered from a certain exhaustion and published little during his early years in Paris, he later recovered his former productivity and retained it almost until his death in 1813 [Delambre, 1814].

Lagrange's whole biography draws a picture of a man with no interest in political events, withdrawn and largely detached from external influences, but an extremely productive mathematician. His extensive *Oeuvre* not only encompasses analysis (theory of ordinary and partial differential equations, calculus of variations, theory of functions) and mathematical physics (potential theory, celestial mechanics, the three-body problem), but also in-

cludes works on algebra, differential geometry, number theory and other branches of mathematics [Taton, 1974]. Lagrange abandoned intuitively geometric considerations, which he systematically avoided in publications such as the *Méchanique analitique*, thereby establishing his reputation as a 'pure analyst'. In the context of discovery, however, he stressed the heuristic value of geometrical intuition [Grattan-Guinness, 1981, 679]. Thus, when he warns with pride in the foreword of his great textbook on mechanics (our concern here) that 'no figures are to be found in this book' (p. vi), this can also be taken with a grain of salt in regard to the origin of the work: it is an important stipulation in regard to the purification of rational mechanics as a science but not in regard to the heuristics he might have used in working it out.

2 LAGRANGE'S CONCEPTION OF 'ANALYTIC MECHANICS' AS A SCIENCE

Lagrange's claim to 'freedom from geometry' is closely connected to his commitment to make mechanics into a 'new branch' of analysis (p. vi). Before turning to the contents of the book, we would like to go into the question of how it came by its title, and in which tradition sought thereby to classify it. At the time of Lagrange, the adjective 'analytic' in mechanics was no longer used in the context of the old scientific distinction between analytic and synthetic methods, as in the *metodo risolutivo* and *metodo compositivo* of Galileo Galilei. Following this distinction, 'synthesis' came essentially to mean the inductive demonstration of *grounds of explanation*, and 'analysis' the deductive derivation of *hypotheses of explanation*. From the time of Euler, on the other hand, the objective 'analytic' within rational mechanics was used mainly or (see below) even entirely to designate the introduction of mathematical methods: analytical mechanics makes use of the (higher) differential and integral calculus or the 'analysis of the infinite'. Mathematical analysis thus lay behind the naming of analytical mechanics in a double sense, first in the formulation and description of its first principles or axioms, and second as an instrument of derivation, that is, the means whereby empirically demonstrable consequences can be deduced from the 'analytical principles'. The latter constitutes the particular meaning of the *calculus of variations* (as a part of higher analysis) for that branch of mathematics. Lagrange's *Méchanique analitique* thus also became the first textbook of mechanics in which the calculus of variations finds extensive application.

When understood in this mathematical sense, analytical mechanics has as its 'synthetic counterpart' *geometrical mechanics*, which was likewise further developed in the 19th century as a part of theoretical mechanics [Ziegler, 1985]. Of course from the middle of the 18th century, the analytic approach, because its generality and the power of its methods, gained a strong predominance, especially in view of the canonized definition of theoretical or 'rational' mechanics by Isaac Newton (1643–1727) in his *Principia* (1687) as an 'exact' and 'established' science ([Newton, 1726, xvii]; compare §5.11).

While Newton used the still prevalent synthetic (in the geometric sense) methods, Euler was the first to point out, in stressing the significance of mathematical analysis for mechanics, that it could not only lead to true mathematical statements about nature but also to a 'sufficiently clear and definite knowledge of itself' [Euler, 1736, vol. 1, 8]. His *Mechanica*

was described by Lagrange as 'the first work [...] in which analysis has been applied to the science of motion' (*Oeuvres*, vol. 11, 243). It is important to observe, however, that geometric methods retained their importance in most of his later works on mechanics used by Euler in his *Mechanica*.

Half a century later Lagrange himself, in the *Méchanique analitique*, was the first explicitly to introduce a monism of 'analytic methodology' into rational mechanics. He claimed that non-analytic methods could be *universally* relinquished, thereby turning mechanics into a 'new branch' of analysis; all 'geometric' and even 'mechanical considerations' would thus become superfluous (p. vi).

An important factor for Lagrange was the *unification* of analytically formulated principles of mechanics: starting from *a single* principle, namely, his reworking of the principle of virtual velocities, which was a 'combination' of the older form of that principle and the principle named after D'Alembert, it should be possible to develop in a deductive way the whole of statics and dynamics (see section 3 below). In this metascientific respect, the *Méchanique analitique* is a consistently 'synthetic' work, in the old (for example, Galilean) sense of the word.

Nevertheless the book differs from the great old 'synthesized' textbooks of mechanics in one important respect. As a prelude to an axiomatic construction modeled in Euclid's *Elements* of mechanics based on an explanation of the conceptual assumptions of that subject, especially the fundamental notions of space, time, matter or mass, and force, the systematic part of the *Méchanique analitique* begins directly with a discussion of analytic principles and their interrelations. Lagrange's methodology implied not only the exclusion of other (non-analytic) mathematical methods, but also of non-mathematical methods. Particularly conspicuous by their absence from his magnum opus are the philosophical and scientific reflections on the foundations of mechanics which had up to that time been included within textbooks, such as those of Newton, Euler, and D'Alembert [Newton, 1726; Euler, 1736; D'Alembert, 1743; see Pulte, 2001]. This was an important precedent, since Lagrange's textbook was to become a paradigm for analytical mechanics in the first half of the 19th century. It was a result of this that, at that time, the subdiscipline of mechanics, although belonging to mathematical physics, was often treated more as a branch of 'pure' mathematics, and thus achieved the status of an 'infallible' mathematical science [Pulte, 2005].

3 LAGRANGE'S FUNDAMENTAL DEVELOPMENTS UP TO THE FIRST EDITION OF THE *MÉCHANIQUE ANALITIQUE*

In the structure of a textbook that claims, in the sense described above, to turn an empirical science like mechanics into a 'purely' mathematical one, the determination of the principle or principles in which the axiomatic-deductive construction is to be based is of the greatest importance. Lagrange's attitude in this respect was by no means uniform, but, in both editions of the *Méchanique analitique*, subject throughout to variations that we shall describe in this section. We shall follow Lagrange's style, referring to the first edition as the *Méchanique analitique*, and to the second edition (1811–1815) as the *Mécanique analytique* (where we cite the slightly revised fourth edition in the *Oeuvres*).

Lagrange's first preference, during his early years at Turin, was for the principle of least action, formulated by P.L.M. de Maupertuis (1698–1759) and Euler, as 'the universal key to all problems both in statics and dynamics' (letter to Euler of 19 May 1756; see *Oeuvres*, vol. 13, 392). In his publications relating to this topic, he first develops his new calculus of variations on the δ-calculus [Lagrange, 1760a] and then gives an application of it in mechanics [Lagrange, 1760b]. For this *application*, he goes directly to the *Additamentum* II of Euler's *Methodus inveniendi* (§12) and gives the following general formulation of the principle of least action [Lagrange, 1760b, 366]:

$$\delta \int u\,ds = \int \delta(u\,ds) = \int (\delta u\,ds + u\delta\,ds) = 0. \tag{1}$$

A mass M with velocity u moves under conservative external forces in such a way that the energy integral achieves an extremum (not necessarily a minimum), that is, according to (1), the first variation vanishes. Lagrange is thus proceeding in the variation process from an iso-energetic variation.

Lagrange generalizes the formulation (1) of the principle of least action not only to problems with n mass-points but also to problems with constraints and finally to the mechanics of solid and fluid continua. The deductive power of the principle is emphasized above all by his derivation from it, using the δ-calculus, of the theorem on the conservation of the motion of the centre of mass and the plane theorem (on the conservation of angular momentum). One of the most important achievements of Lagrange's *Application* is the derivation [Lagrange, 1760b, 369] from (1) of the so-called Newtonian equations of motion for a mass subject to conservative central forces (with Cartesian coordinates (Π, Ω, Ψ)):

$$d\frac{u\,dx}{ds} + \Pi\,dt = 0, \qquad d\frac{u\,dy}{ds} + \Omega\,dt = 0, \qquad d\frac{u\,dz}{ds} + \Psi\,dt = 0. \tag{2}$$

It is particularly noteworthy that Lagrange thereby goes beyond Euler's *narrow form* of the principle of least action and considers an application to a system in which conservation of active force is not generated [Lagrange, 1760b, 384–385]. He therefore takes into consideration a *wider form* of the principle of least action, a form that Maupertuis had in mind but was not able to specify. In the applications, however, Lagrange always restricts himself to *conservative* systems, in which the conservation of living force is guaranteed.

In his 'Researches on the libration of the Moon' [Lagrange, 1764] he departs for the first time from his attempts to axiomatize mechanics on the basis of the principle of least action, and replaces the principle of Euler and Maupertuis by a generalization of D'Alembert's principle. Later, in his book, he describes this generalization as the 'principle of virtual velocities' (p. 12; compare p. 8). He never explained this change of heart, although the final unclear relationship between the conservation of *vis viva* and the principle of least action might be seen as a major reason for this: Lagrange is never clear as to whether this conservation law in his formulation of the energy principle should be regarded as a *hypothesis* or (as desired) as a *result*. A second reason could be that he regarded the principle of virtual velocities as fundamental in the domain of statics [Lagrange, 1764, 10], and so it seemed to him that a uniform axiomatization of statics and dynamics might be easier to achieve on the basis of that principle [Pulte, 1989, 252–261].

There is support for both reasons in the *Méchanique analitique*. It is divided into two parts, statics (pp. 1–157) and dynamics (pp. 158–512), and Lagrange sets out to base both parts solely on the principal of virtual velocities, paying particular attention in the second part to the derivation of vis viva-conservation (pp. 205–208) and annexing that of the principle of least action (pp. 208–212).

Lagrange introduces the principle of virtual velocities in the first edition as 'a kind of axiom for mechanics' (p. 12) for *statics*, where it 'has all the simplicity one might desire in a fundamental principle' (p. 10). By statics he means the 'science of equilibrium of forces' (p. 1), as he says right at the beginning. If one now considers a system of mass-points in a static equilibrium acted on at any given time by forces $P, Q, R, \ldots$ and gives it a small perturbation, then the individual masses experience 'virtual' displacements, that is, displacements compatible with any connections that may exist between the masses. Let $\delta p, \delta q, \delta r, \ldots$ be their projections on the forces $P, Q, R, \ldots$, with the sense of direction of the projection indicated by a suitable choice of sign. Lagrange labels these displacements as 'virtual velocities' by appealing to a fixed time element dt. The *principle of virtual velocities* (or *displacements*) now asserts that a system is in equilibrium if the sum of the 'moments of force' vanishes (p. 15):

$$P\delta p + Q\delta q + R\delta r + \cdots = 0. \tag{3}$$

He then applies this relation, from 'Section III' of the *Méchanique analitique*, in the treatment of general properties of the equilibrium of point systems (Section III), methods for solving the resulting equations (Section IV), special problems in statics (Section V), hydrostatics (Section VI), problems of equilibrium of incompressible fluids (Section VII) and problems of equilibrium of compressible and elastic fluids (Section VIII).

Lagrange constructs *dynamics* in an entirely analogous way (see section 4 below). He first extends, as in the earlier *Recherches* [Lagrange, 1764, 10], the principle of virtual velocities to problems of motion in that, as well as the external forces $P, Q, R, \ldots$, he also takes into account on the individual point masses their accelerations, which must be compatible with the connections within the system. Multiplication by the instantaneous masses yields the forces that the same accelerations would produce in free masses. His claim is then that under a virtual displacement the 'moments of the forces' $P, Q, R, \ldots$ must be equal to the moments of these forces of acceleration, where the sign difference depends on a convention (today reversed) on the direction of action of the $P, Q, R, \ldots$ (p. 195):

$$m\left(\frac{d^2x}{dt^2}\delta x + \frac{d^2y}{dt^2}\delta y + \frac{d^2z}{dt^2}\delta z\right) = -m(P\delta p + Q\delta q + R\delta r + \cdots). \tag{4}$$

If we depart from Lagrange's sign convention and write the central forces $P, Q, R, \ldots$ in Cartesian coordinates, then Lagrange's general fundamental law of dynamics takes the following slightly 'modernized' form (compare p. 200):

$$\sum_{i=1}^{n}\left[\left(X_i - m_i\frac{d^2x_i}{dt^2}\right)\delta x_i + \left(Y_i - m_i\frac{d^2y_i}{dt^2}\right)\delta y_i + \left(Z_i - m_i\frac{d^2z_i}{dt^2}\right)\delta z_i\right] = 0. \tag{5}$$

If n mass-points m_i with coordinates (x_i, y_i, z_i) are subjected at any given moment to forces with Cartesian coordinates (X_i, Y_i, Z_i) and each mass-point undergoes a virtual displacement $(\delta x_i, \delta y_i, \delta z_i)$, then the sum of the 'moments' vanishes according to (5), as does the 'virtual work' of the forces (X_i, Y_i, Z_i) reduced by the forces of inertia. In the case of 'free' systems, the principle is really trivial: the displacements $(\delta x_i, \delta y_i, \delta z_i)$ can then be chosen arbitrarily, so that the summands in (5) must vanish individually, which easily yields the differential equation of motion:

$$X_i = m_i \frac{d^2 x_i}{dt^2}, \quad Y_i = m_i \frac{d^2 y_i}{dt^2}, \quad Z_i = m_i \frac{d^2 z_i}{dt^2} \quad (1 \leq i \leq n). \tag{6}$$

In the case of equilibrium, (5) takes the form

$$\sum_{i=1}^{n} [X_i \delta x_i + Y_i \delta y_i + Z_i \delta z_i] = 0; \tag{7}$$

or, if one starts with forces $P, Q, R, \ldots$ that are not written in Cartesian coordinates but act along any lines $p, q, r, \ldots$, one obtains the original Lagrangian formulation (3) for statics (p. 15):

$$P\delta p + Q\delta q + R\delta r + \cdots = 0. \tag{8}$$

This formula clearly expresses the vanishing of the sum of the moments in the equilibrium case.

The coordinates (x_i, y_i, z_i) of the mass-points are subject to m equations (assumed independent) altogether (p. 227):

$$L_j(x_i, y_i, z_i) = 0 \quad (1 \leq j \leq m, m < 3n). \tag{9}$$

In this case of holonomous constraints (that is, constraints which can be given in form of an equation) or even skleronomous constraints (i.e. constraints with time-independent equations), the displacements $(\delta x_i, \delta y_i, \delta z_i)$ in (5) are no longer arbitrary but chosen in accordance with (9). According to the method of the Lagrange multiplier, the differential equations (6) of the free motion are replaced by the following equations of motion, which are today usually referred to as 'Lagrangian equations of the first kind' (p. 228):

$$m_i \ddot{x}_i = X_i + \sum_{k=1}^{m} \lambda_k \frac{\partial L_k}{\partial x_i}, \; m_i \ddot{y}_i = Y_i + \sum_{k=1}^{m} \lambda_k \frac{\partial L_k}{\partial y_i}, \; m_i \ddot{z}_i = Z_i + \sum_{k=1}^{m} \lambda_k \frac{\partial L_k}{\partial z_i}. \tag{10}$$

The expressions under the summation sign in (10) can be interpreted physically as the constraining forces on the masses m_i needed to ensure the fulfillment of the conditions (9). The fact that these equations, from which the multiplicands λ_k must be eliminated (as is always possible under the above assumption), uniquely determine the (x_i, y_i, z_i) using the equations (9), was first shown explicitly much later by C.G.J. Jacobi (1804–1851) [Jacobi, 1884, 132–133].

Lagrange himself followed a rather different route. Given the m independent conditions (9) on the $3n$ time-dependent coordinates (x_i, y_i, z_i), one can define $(3n - m)$ independent

variables $q_j = q_j(t)$ (today called 'Lagrangian' or 'generalized coordinates') in such a way that the equations (9) hold identically and no further constraints are introduced. Using the so-called 'generalized forces' Q_r given by

$$\sum_{i=1}^{n}\left[X_i\frac{\partial x_i}{\partial q_r}+Y_i\frac{\partial y_i}{\partial q_r}+Z_i\frac{\partial z_i}{\partial q_r}\right]=\frac{dU}{dq_r}=:Q_r, \tag{11}$$

where U denotes the *potential function* (which exists in our case), and the further abbreviation

$$T:=\frac{1}{2}\sum_{i=1}^{n}m_i\left(\dot{x}_i^2+\dot{y}_i^2+\dot{z}_i^2\right) \tag{12}$$

for the total *kinetic energy* of the mechanical system, one obtains in place of (10) the following 'Lagrangian equations of the second kind' (p. 226):

$$\frac{d}{dt}\left(\frac{\partial T}{\partial\dot{q}_r}\right)-\frac{\partial T}{\partial q_r}=Q_r\quad(1\le r\le 3n-m). \tag{13}$$

Here $\dot{q}_r = dq_r/dt$ is the 'generalized velocity'.

The derivation of (13) from (5) is one of the most significant achievements of the *Méchanique analitique*, since all dynamical problems satisfying the given conditions can be reduced with the aid of (13) to the determination of the two functions kinetic energy and potential energy (confer T and U above).

From the point of view of the formal and deductive organization of mechanics, it seems, generally speaking, entirely plausible that Lagrange described the principle of virtual velocities in the *statics* part of his work as 'a kind of axiom for mechanics' (collectively) because it also demonstrates its power in *dynamics*. In Section III of the dynamics part, Lagrange deduces from (10) the law of conservation of motion of the centre of mass (p. 201), the law of areas or conservation of angular momentum (p. 205), at least in the case when the mechanical system is subject to the time-independent constraints (9) and only governed by central forces, and also the law of conservation of active force (p. 208) and the principle of least action. The latter struck him as a 'very remarkable property of motion' (p. 211), but was without a teleological meaning for him (p. 196; see Pulte [1989, 252–261]).

As to the 'principle' aspect of the *Méchanique analitique*, we only note here that Lagrange's application of the principle of virtual velocities and its consequences such as (13) in subsequent Sections V–X of the dynamics part of his work throw into bold relief the tension between his formal-abstract approach to mechanics and physical intuition, especially in the treatment of the mechanics of solid continua and hydrodynamics [Grattan-Guinness, 1990a, vol. 1, 286–287]. This tension, even at the level of a discussion of principles (see section 5 below), also plays an important role in the reception of the work. Before turning to this, we give a summary of the contents of the whole work.

4 CONTENTS OF THE EDITIONS

Table 1 shows the contents of the book, and also of the later editions. Under the heading 'topic', the original headings of the individual Sections are translated as literally as possible in order to highlight the differences between the first edition of 1788 and the second edition of 1811–1815. The Section headings and contents of the second, third and fourth editions are grouped together, with the pagination referring to the two-volume second edition.

Table 1. Summary by Sections of the two editions of Lagrange's book.

	Méchanique Analitique (1788)		***Mécanique Analytique*** (2nd–4th editions)	**2nd ed.**
Sec.	**Topic**	**p.**	**Topic**	**p.**
	Part 1: *Statics.*		Part 1: *Statics.*	
I	On the different principles of statics.	1	On the different principles of statics.	Vol. 1, 1
II	General formula for the equilibrium of any system of forces; with a method of using it.	12	General formula of statics for the equilibrium of any system of forces, with a method of using it.	27
III	General properties of equilibrium deduced from the preceding formula.	25	General properties of equilibrium of a system of bodies, deduced from the preceding formula.	45
IV	Very simple method of finding the necessary equations of equilibrium for any system of bodies regarded as points, or as finite masses, and underlying given forces.	44	A more general and simpler way to use the formula of equilibrium, demonstrated in Section II.	77
V	Solution of different problems of statics.	58	Solution of different problems of statics.	113
VI	On the principles of hydrostatics.	122	On the principles of hydrostatics.	189
VII	On the equilibrium of incompressible fluids.	130	On the equilibrium of incompressible fluids.	197
VIII	On the equilibrium of compressible and elastic fluids.	155	On the equilibrium of compressible and elastic fluids.	231

Table 1. (*Continued*)

	Part 2: *Dynamics.*		Part 2: *Dynamics.*	
I	On the different principles of dynamics.	158	On the different principles of dynamics.	237
II	General formula for the motion of a system of bodies animated by any forces.	189	General formula of dynamics for the motion of a system of bodies animated by any forces.	263
III	General properties of motion deduced from the preceding formula.	198	General properties of motion deduced from the preceding formula.	273
IV	A simpler method for arriving at the equations which determine the motion of any system animated by any accelerating forces.	216	Differential equations for the solution of all problems of dynamics.	325
V	Solution of different problems of dynamics.	233	General method of approximation for the problems of dynamics, based on the variation of arbitrary constants.	345
VI	On the rotation of bodies.	337	On the very small oscillations of any system of bodies. [End 422.]	369
VII	On the principles of hydrodynamics.	428	On the motion of a system of free bodies treated as mass points and acted upon by forces of attraction.	Vol. 2, 1
VIII	On the motion of incompressible fluids.	437	On the motion of bodies which are not free, and which interact in an arbitrary manner.	177
IX	On the motion of compressible and elastic fluids. [End 512.]	492	On rotational motion. On the rotation of an arbitrary system of bodies.	211
X			On the principles of hydrodynamics.	277
XI			On the motion of incompressible fluids.	286
XII			On the motion of compressible and elastic fluids.	337
Note I			On the determination of the orbits of planets.	355
Note II			On rotational motion.	357
			List of the works of M. Lagrange. [End 378.]	372

5 FUNDAMENTAL DIFFERENCES BETWEEN THE *MÉCHANIQUE ANALITIQUE* AND THE *MÉCANIQUE ANALYTIQUE* AND LATER REVIEWS

Table 1 makes it clear that there are significant differences between the first edition of Lagrange's magnum opus on the one hand and the second (and subsequent) editions on the other, especially in the treatment of *dynamics* in Sections IV–XII. These alterations and their background cannot be studied in detail here. Of continuing importance for us, of course, is how Lagrange seeks to justify his commitment to pass from 'self-evident' analytical principles of mechanics to a 'purely analytic' foundation of that subject. In this context, it is of interest to confirm that all the important extensions of the principle of virtual velocities, as adopted by Lagrange, up to round 1850 take as their starting point Lagrange's formulation (5), or (7) or (8) in the case of statics.

A first important point for discussion was the *axiomatic status* of the principle. In his later years, Lagrange himself no longer saw it sufficiently self-evident to be adopted as an axiom, mainly because of a critical analysis by J.-B. J. Fourier (1768–1830) of Lagrange's formulation of the principle [Fourier, 1798; see Costabel, 1972]. This is clearly expressed in the second edition (*Mécanique analytique*, vol. 1, 23), where Lagrange gives two different 'proofs' (so called by him) to compensate for this. The first was an attempt to provide an incontestable self-evident justification for the static principle by appealing to simple machines such as the pulley [Lagrange, 1798 and *Mécanique analytique*, vol. 1, 18–26; see Pulte, 1998, 165–166]. The second was undertaken shortly before his death, in the new edition of the *Théorie des fonctions analytiques*, again with the intention of passing in this way from statics to dynamics [Lagrange, 1813, 350–357 and 1847, 377–385; see Pulte, 1998, 169–172] and thereby rescuing his programme of putting mechanics on a 'purely analytic' basis. This second attempt does not appear again in the *Mécanique analytique*.

Other attempts at proof that concern Lagrange's magnum opus but cannot be pursued here [see Lindt, 1904; Grattan-Guinness, 1990a, vol. 1, 302–308; Benvenuto, 1991, vol. 1, 95–115]. They were undertaken especially by A.M. Ampère (1775–1836) in his [1806], L.N.M. Carnot (1753–1823) in [1803], A.A. Cournot (1801–1877) in [1827], J. Fossombroni (1754–1844) in [1796], [Laplace, 1799], L. Poinsot (1777–1859) in [1806, 1838, 1846], S.D. Poisson (1781–1840) in [1833], and G.C.F.M. de Prony (1755–1839) in [1798]. The main purpose was to reduce the principle of virtual velocities to the lever principle or the parallelogram law of forces.

The idea, rather strange by today's standards, of a mechanical principle or axiom that stands at the head of a deductive system but cannot itself be deduced, opens up a 'crisis of principles' [Bailhache, 1975, 7] that ultimately has its roots in the fact that viewing analytical mechanics in the tradition of Lagrange according to the pure mathematical ideals of reason and exactness is a misconception that cannot be resolved at the level of principles or axioms. Jacobi was the first, with his *Vorlesungen über analytische Mechanik* of 1847–1848 [Jacobi, 1996, 29–39, 59–94; see Pulte, 1998], to make a clear break with this 'certistic' tradition and, in the second half of the century, came increasingly to favour the opinion that such a principle cannot be secured either mathematically or metaphysically, but rather must be regarded as a hypothetical assumption.

A *second* point closely related to the problem of *justifying* the principle concerns its *generalization* to non-holonomic and non-skleronomic systems: Fourier had already pointed out that Lagrange's principle, even in the static form (8), fails when the constraints are given in the form of an inequality rather than an equation (9) [Fourier, 1798, 30], as in the case of a mass-point moving *outside* a surface rather than *on* it. (9) must then be replaced by

$$L_j(x_i, y_i, z_i) \leq 0 \quad (1 \leq j \leq m,\ m < 3n). \tag{14}$$

The principle of virtual velocities in Lagrange's form (8) then generalizes to the assertion that the total moment or virtual work in the equilibrium case does not vanish but (with a suitable sign convention) becomes negative (cf. Jacobi [1996, 40], for example):

$$P\delta p + Q\delta q + R\delta r + \cdots \leq 0. \tag{15}$$

A corresponding generalization of the principle for *statics* was proposed, to some extent without awareness of Fourier's criticisms and related investigations, in [Cournot, 1827; Gauss, 1829] and [Ostrogradsky, 1838a]; the first extension to dynamics seems to be due to [Ostrogradsky, 1838b].

Another desirable generalization was pointed out by Lagrange himself in the first edition of the *Méchanique analitique* (p. 198) but later passed over in silence, and this is the extension of the principle to *rheonomic* systems, where the equations (5) depend on the time t:

$$L_j(x_i, y_i, z_i, t) = 0 \quad (1 \leq j \leq m,\ m < 3n). \tag{16}$$

Poisson argued that the application of the principle (1) in this case introduced only infinitesimally small errors and was thus legitimate [Poisson, 1833, vol. 2, arts. 564–573; see Lindt, 1904, 170f.]; however, Ostrogradsky [1842] advocated the view that in this case the time should be regarded as a variable and thus (1) should be supplemented by a variation in the time. It turned out that this is not necessary, and (1) can also be applied in the case of rheonomic constraint [Schaefer, 1919, 13–14]. But this problem was not effectively cleared up until the end of the 19th century—compare [Hertz, 1894] (§52), [Voss, 1901] and [Schaefer, 1919], for example—along with the extension of the principle to problems in which, as in the case of friction, no force function as in (11) can be given or in which there are non-holonomic conditions connecting not the coordinates of the mass-points themselves, but those of their first derivatives, in the form of an equation (analogous to (9)), as in the case of a rolling motion without slipping, say of a sphere on a horizontal plane.

A *third* point in the century-long discussion of the principle of virtual velocities concerns its *physical relevance*, and thus ultimately that of the whole of analytical mechanics in the sense of Lagrange. In the tradition of French mathematical physics following Lagrange, there appeared a 'physicalization' of mechanics, which found an outlet with Poisson in the explicit movement away from Lagrangian and traditional *analytical* mechanics towards a specific and new *physical* mechanics [Duhem, 1980; Grattan-Guinness, 1990a, vol. 2].

One of the important questions here concerned the nature of the constraints required to make the moving mass-points satisfy quasi-geometric restrictions of the form (9)

[Duhem, 1980, 37–46]. This question was later taken up by Jacobi, among others, and applied to Lagrange's passage from statics to dynamics via the principle of virtual velocities. While in the equilibrium case the constraints only have to compensate for the forces such as weight on the masses at rest, in the case of motion (as Jacobi, unlike Lagrange, proved by exclusively analytic methods) they also depend on the velocities and they do not in general satisfy the assumptions made by Lagrange. Finally, in his analytical mechanics Lagrange had blended forces (such as gravitation) with something 'entirely heterogeneous', namely, pure mathematical conditions of the form (9), and this made it impossible to regard the general dynamical form of the principle (5) as an extension of the static principles (7) or (8) as Lagrange claimed [Jacobi, 1996, 85–88]. Jacobi's main criticism was that Lagrange's use of this principle totally ignored the physical constitution of the body in favour of the 'purity' of an analytic mechanics which even waived its own justification in order to remain 'pure' [Jacobi, 1996, 193–194]. From the middle of the 19th century, this and related criticisms led to investing the principal of virtual velocities with a stronger physical relevance that replaced conditions of the Lagrangian type (9) by external forces with 'more realistic' laws of force or potential and used them in applications of (5).

6 GENERAL ASSESSMENT AND RECEPTION OF THE WORK

In this sketch we have tried to describe Lagrange's discussion of principles and the nature and reception of the *Méchanique analitique* in that context. Of course the significance and influence go far beyond (confer [Cayley, 1857]). It was certainly regarded as the most important unification of rational mechanics at the turn of the 18th century and as its 'crowning' [Dugas, 1955, 332]. This achievement of unification and the abstract-formal nature of the work, physically reflected in immediate applications, earned the extravagant praise of Ernst Mach: 'Lagrange [...] strove to dispose of all necessary considerations *once and for all*, including as many as possible in one formula. Every case that arises can be dealt with according to a very simple, symmetric and clearly arranged scheme [...] Lagrangian mechanics is a magnificent achievement in respect of the economy of thought' ([Mach, 1933, 445]; see [Fraser, 1990] concerning Auguste Comte). We make a marginal note that Mach's own history of mechanics participated strongly in the historical fulfillment which Lagrange placed at the beginning of his work (pp. 1–12) and carried much further in the second edition (*Mécanique analytique*, vol. 1, 1–26).

But the price Lagrange had to pay for this 'magnificent achievement' must not be overlooked. For all its mathematical elegance, described by W.R. Hamilton as 'a kind of scientific poem' [Hamilton, 1834, 134], the *Méchanique analitique* is a physically and philosophically *sterile* work, and necessarily so because it claims to be a 'purely analytical' mechanics. Lagrange contributed very little towards the conceptual development of theoretical mechanics and its philosophical foundations [Grattan-Guinness, 1990a, vol. 1, 274–301; Truesdell, 1960].

Lagrange's attempt to 'reduce' mechanics to analysis and thereby to an algebra of power series strikes us today as a misplaced endeavour to mathematicize almost entirely an empirical science, and thus to endow it with infallibility, which does not stand up to critical examination. In this sense Jacobi wrote of the *Méchanique analitique* as follows: 'because

of the significance and authority of the book, one can be led to accept as true and rigorous things that are not [...] I have had pupils who understood the analytical mechanics better than I, but understanding something is not always a good sign' [Jacobi, 1996, 29].

A 'good sign', however, is that a mathematical work can still be read with profit more than 200 years after its appearance. The mathematician will appreciate the elegance and methods of the work while the 'mechanicist' will admire the achievements of systematization and unification. In that sense the *Méchanique analitique* can lay claim today to be 'one of the outstanding landmarks in the history of both mathematics and mechanics' [Sarton, 1944, 470].

BIBLIOGRAPHY

D'Alembert, J. le R. 1743. *Traité de dynamique*, Paris: David. [See §11.]

Ampère, A.M. 1806. 'Démonstration générale du principe des vitesses virtuelles, dégagée de la considération des infiniment petits', *Journal de l'École Polytechnique, (1) 6, cah. 13*, 247–269.

Bailhache, P. 1975. 'Introduction et commentaire', in [Poinsot, 1975], 1–199.

Barroso-Filho, W. 1994. *La mécanique de Lagrange. Principes et méthodes*, Paris: Édition Karthala.

Benvenuto, E. 1991. *An introduction to the history of structural mechanics*, 2 vols., Berlin: Springer.

Borgato, M.T. and Pepe, L. 1987. 'Lagrange a Torino (1750–1766) e le sue lezione inedite nella Reale Scuole di Artiglieria', *Bollettino di storia delle scienze matematiche, 7*, 3–200.

Bos, H.J.M. 1980. 'Mathematics and rational mechanics', in G.S. Rousseau and R. Porter (eds.), *The ferment of knowledge: studies in the historiography of eighteenth-century science*, Cambridge: Cambridge University Press, 327–355.

Carnot, L.N.M. 1803. *Principes fondamentaux de l'équilibre et du mouvement*, Paris: Deterville.

Cayley, A. 1857. 'Report on the recent progress of the theoretical dynamics', *Reports of the British Association of the Advancement of Science*, 1–42.

Costabel, P. 1972. 'Fourier et le principe des vitesses virtuelles', *Sciences, 3*, 235–238.

Cournot, A.A. 1827. 'Extension du principe des vitesses virtuelles au cas où les conditions des liaisons sont exprimées par inégalités', *Bulletin universel des sciences et de l'industrie, sciences mathématiques, 8*, 165–170.

Delambre, J.B.J. 1814. 'Notice sur la vie et les ouvrages de M. le Comte J.L. Lagrange', *Mémoires de la classe des sciences mathématiques de l'Institut*, (1812: publ. 1816), pt. 2, xxxiv–lxxxviii. [Repr. in *Works de Lagrange*, vol. 1, Paris: Gauthier–Villars, 1867, ix–li.]

Dugas, R. 1955. *A history of mechanics* (trans. J.R. Maddox), New York: Central.

Duhem, P. 1980. *The evolution of mechanics* (trans. M. Cole), Alphen an den Rijn and Germantown: Sijthoff & Noordhoff.

Euler, L. 1736. *Mechanica sive motus scientia analytice exposita*, 2 vols., Saint Petersburg: Academy. [Repr. as *Opera omnia*, ser. 2, vols. 1–2, Leipzig: Teubner, 1912.]

Fossombroni, V. 1796. *Memoria sul principio delle velocità virtuali*, Firenze: Cambiagi.

Fourier, J.-B.J. 1798. 'Mémoire sur la statique, contenant la démonstration du principe des vitesses virtuelles, et la théorie des momens', *Journal de l'École Polytechnique, (1) 2, cah. 5*, 20–60. [Repr. in *Œuvres*, vol. 2, Paris: Gauthier–Villars, 1890, 475–521.] [See §26.]

Fraser, C. 1983. 'J.L. Lagrange's early contributions to the principles and methods of mechanics', *Archive for history of exact sciences, 28*, 197–241.

Fraser, C. 1990. 'Lagrange's analytical mechanics. Its Cartesian origins and reception in Comte's positive philosophy', *Studies in the history and philosophy of science, 21*, 243–256.

Gauss, C.F. 1829. 'Über ein neues und allgemeines Grundgesetz der Mechanik', *Journal für die reine und angewandte Mathematik, 4*, 232–235. [Repr. in *Werke*, vol. 5, Leipzig: Teubner, 1868, 23–28.]

Grattan-Guinness, I. 1981. 'Recent researches in French mathematical physics of the early 19th century', *Annals of science, 38*, 663–690.

Grattan-Guinness, I. 1986. 'How it means: mathematical theories in physical theories. With examples from French mathematical physics of the early 19th century', *Rendiconti dell' Accademia Nazionale delle Scienze detta dei XL, (5) 9*, 89–119.

Grattan-Guinness, I. 1990a. *Convolutions in French mathematics, 1800–1840*, 3 vols., Basel: Birkhäuser.

Grattan-Guinness, I. 1990b. 'The varieties of mechanics by 1800', *Historia mathematica, 17*, 313–338.

Hamilton, W.R. 1834. 'On a general method in dynamics', *Philosophical Transactions of the Royal Society*, 247–308. [Repr. in *Mathematical papers*, vol. 2, Cambridge: Cambridge University Press, 1940, 103–161.]

Hertz, H. 1894. *Die Prinzipien der Mechanik in neuem Zusammenhang dargestellt*. Leipzig: Barth. [See §52.]

Itard, J. 1973. 'Lagrange, Joseph Louis', in *Dictionary of scientific biography*, vol. 7, New York: Scribners, 559–573.

Jacobi, C.G.J. 1884. *Vorlesungen über Dynamik. Gehalten an der Universität Königsberg im Wintersemester 1842–1843* (ed. A. Clebsch), Berlin: Springer, 1866. [2nd ed. as suppl. vol. of *Gesammelte Werke*, Berlin: Reimer, 1881.]

Jacobi, C.G.J. 1996. *Vorlesungen über analytische Mechanik. Berlin 1847/48* (ed. H. Pulte), Braunschweig and Wiesbaden: Vieweg.

Jouguet, E. 1908–1909. *Lectures de mécanique*, 2 vols., Paris: Gauthier–Villars.

Jourdain, P.E.B. 1908. (Ed.), *Abhandlungen über die Prinzipien der Mechanik von Lagrange, Rodrigues, Jacobi und Gauss*, Leipzig: Engelmann (*Ostwalds Klassiker der exakten Wissenschaften*, no. 167).

Lagrange, J.L. *Works. Oeuvres*, 14 vols., Paris: Gauthier–Villars, 1867–1892.

Lagrange, J.L. 1759. 'Recherches sur la méthode de maximis, et minimis', *Miscellanea Taurinensia, 1*, 18–32. [Repr. in *Works*, vol. 1, 3–20.]

Lagrange, J.L. 1760a. 'Essai d'un nouvelle méthode pour déterminer les maxima et les minima des formules indéfinies', *Miscellanea Taurinensia, 2* (1760/1761: publ. 1762), 173–195. [Repr. in *Works*, vol. 1, 335–362.]

Lagrange, J.L. 1760b. 'Application de la méthode exposée dans le mémoire précédent à la solution de différens problèmes de dynamique', *Miscellanea Taurinensia, 2* (1760/1761: publ. 1762), 196–268. [Repr. in *Works*, vol. 1, 365–468.]

Lagrange, J.L. 1764. 'Recherches sur la libration de la Lune dans lesquelles on tâche de résoudre la question proposés par l'Académie Royale des Sciences, pour le Prix de l'année 1764', *Prix de l'Académie Royale des Sciences de Paris, 9* (1777), 1–50. [Repr. in *Works*, vol. 6, 5–61.]

Lagrange, J.L. 1770. 'Sur la méthode des variations', *Miscellanea Taurinensia, 6* (1766–1769: publ. 1773), 163–187. [Repr. in *Works*, vol. 2, 37–63.]

Lagrange, J.L. 1797. *Théorie des fonctions analytiques*, 1st ed., Paris: L'Imprimerie de la République. [See §19.]

Lagrange, J.L. 1798. 'Sur le principe des vitesses virtuelles', *Journal de l'École Polytechnique, (1) 2, cah. 5*, 115–118. [Repr. in *Works*, vol. 3, 315-321.]

Lagrange, J.L. 1813. 'Théorie des fonctions analytiques', *Journal de l'École Polytechnique, (1) 3, cah. 9*, 1–383. Also as *Théorie des fonctions analytiques*, 2nd ed., Paris: Bachelier.

Lagrange, J.L. 1847. *Théorie des fonctions analytiques*, 3rd ed., Paris: Bachelier. [Repr. as *Works*, vol. 9.]

Laplace, P.S. 1799. *Traité de mécanique céleste*, vol. 1, Paris: Duprat. [See §24.]

Lindt, R. 1904. 'Das Prinzip der virtuellen Geschwindigkeiten: Seine Beweise und die Unmöglichkeit seiner Umkehrung bei Verwendung des Begriffes "Gleichgewicht eines Massensystems" ', *Abhandlungen zur Geschichte der mathematischen Wissenschaften, 18*, 145–196.

Loria, G. 1949. 'Essai d'une bibliographie de Lagrange', *Isis, 40*, 112–117.

Mach, E. 1933. *Die Mechanik in ihrer Entwicklung, historisch-kritisch dargestellt*, 9th ed., Leipzig: Brockhaus. [Repr. Darmstadt: Wissenschaftliche Buchgesellschaft, 1982.]

Newton, I. 1726. *Principia mathematica*, 3rd ed., London: Innys. [See §5.]

Ostrogradsky, M.W. 1838a. 'Considérations générales sur les moments des forces', *Mémoires de l'Académie des Sciences de St. Pétersbourg, sciences mathematiques et physiques, (6) 1*, 129–150.

Ostrogradsky, M.W. 1838b. 'Mémoire sur les déplacements instantanés des systèmes assujettis à des conditions variables', *Ibidem*, 565–600.

Ostrogradsky, M.W. 1842. 'Sur le principe des vitesses virtuelles et sur la force d'inertie', *Bulletin scientifique publié par l'Académie des Sciences de St. Pétersbourg, 10*, 34–41.

Poinsot, L. 1806. 'Théorie générale de l'équilibre et du mouvement des systèmes', *Journal de l'École Polytechnique, (1) 6, cah. 13*, 206–241.

Poinsot, L. 1838. 'Note sur une certaine démonstration du principe des vitesses virtuelles, qu'on trouve au chapitre III de livre 1er de la "Mécanique céleste" ', *Journal de mathématiques pures et appliquées, (1) 3*, 244–248.

Poinsot, L. 1846. 'Remarque sur un point fondamental de la Mécanique analytique de Lagrange', *Journal de mathématiques pures et appliquées, (1) 11*, 241–253.

Poinsot, L. 1975. *La théorie générale de l'équilibre et du mouvement des système* (ed. P. Bailhache), Paris: Vrin.

Poisson, S.D. 1833. *Traité de mécanique*, 2nd ed., 2 vols., Paris: Bachelier.

De Prony, G.C.F.M. 1798. 'Sur le principe des vitesses virtuelles', *Journal de l'École Polytechnique, (1) 2, cah. 5*, 191–208.

Pulte, H. 1989. *Das Prinzip der kleinsten Wirkung und die Kraftkonzeptionen der rationalen Mechanik. Eine Untersuchung zur Grundlegungsproblematik bei L. Euler, P.L.M. de Maupertuis und J.L. Lagrange*, Stuttgart: Steiner.

Pulte, H. 1998. 'Jacobi's criticism of Lagrange: the changing role of mathematics in the foundations of classical mechanics', *Historia mathematica, 25*, 154–184.

Pulte, H. 2000. 'Beyond the edge of certainty: reflections on the rise of physical conventionalism in the 19th century', *Philosophiae scientiae. Travaux d'histoire et de philosophie des sciences, 4*, 47–68.

Pulte, H. 2001. 'Order of nature and orders of science. mathematical philosophy of nature from Newton and Euler to Kant and Lagrange: some observations and reflections on changing concepts of science', in W. Lefèvre (ed.), *Between Leibniz, Newton, and Kant. Philosophy of science in the eighteenth century*, Dordrecht: Kluwer, 61–92.

Pulte, H. 2005. *Axiomatik und Empirie. Eine wissenschaftstheoriegeschichtliche Untersuchung zur mathematischen Naturphilosophie von Newton bis Neumann*, Darmstadt: Wissenschaftliche Buchgesellschaft.

Sarton, G. 1944. 'Lagrange's personality (1736–1813)', *Proceedings of the American Philosophical Society, 88*, 456–496.

Schaefer, C. 1919. *Die Prinzipe der Dynamik*, Berlin and Leipzig: De Gruyter.

Szabó, I. 1979. *Geschichte der mechanischen Prinzipien und ihrer wichtigsten Anwendungen*, Basel: Birkhäuser.

Taton, R. 1974. 'Inventaire chronologique de l'oeuvre de Lagrange', *Revue d'histoire des sciences et de leurs applications, 27*, 3–36.

Truesdell, C.A. 1960. 'A program toward rediscovering the rational mechanics of the age of reason', *Archive for history of exact sciences, 1*, 1–36.

Voss, A. 1901. 'Die Prinzipien der rationellen Mechanik', in *Encyclopädie der mathematischen Wissenschaften mit Einschluß ihrer Anwendungen*, vol. 4, pt. 1, Leipzig: Teubner, 1–121.

Ziegler, R. 1985. *Die Geschichte der geometrischen Mechanik im 19. Jahrhundert*, Stuttgart: Steiner.

CHAPTER 17

GASPARD MONGE, *GÉOMÉTRIE DESCRIPTIVE*, FIRST EDITION (1795)

Joël Sakarovitch

On the one hand, descriptive geometry is the culmination of a long and slow evolution of different graphical methods used for representing space. On the other hand, it is the fruit of the fertile imagination of a talented geometrician, heir to the age of enlightenment, committed revolutionary, and brilliant teacher. This ambiguous status between art and science undoubtedly confers to descriptive geometry both its charm and specificity. And if the first goal of Monge was a technical one, Michel Chasles was to consider that the 'New geometry' was born with the Monge lectures.

First publication. In *Les Séances des écoles normales recueilliées par des sténographes et revues par des professeurs*, Paris, vol. 1, 49–64 (lecture on 1 pluviose Year III, 1795), 278–285 (9 pluviose), 401–413 (21 pluviose); vol. 2, 149–171 (1 ventose), 338–368 (11 ventose); vol. 3, 61–106 (21 ventose), 332–356 (1 germinal); vol. 4, 87–99 (11 germinal), 291–313 (21 germinal); vol. 7, debates, 28–34 (11 pluviose), 63–74 (16 pluviose), 144–151 (26 pluviose).

Later editions. All Paris. 2nd (ed. J.N.P. Hachette), Baudouin, 1799. 132 pages. 3rd (ed. and suppl. Hachette), Courcier, 1811. 4th (ed. B. Brisson), Courcier, 1820. Further reprs., inc. in 2 vols., Gauthier–Villars, 1922.

New edition (cited here). In *L'Ecole normale de l'an III, Leçons de mathématiques, Laplace, Lagrange, Monge* (ed. J. Dhombres), Paris: Dunod, 1992, 267–459.

English translation of the 2nd ed. Descriptive geometry (trans. F.J. Heather), London: Lockwood, 1809. [Various reprs.]

Spanish translation of the 2nd ed. Geometria descriptiva, Madrid: Impreso Real, 1803.

German translations. 1) *Lehrbuch der darstellende Geometrie* (trans. G. Schreiber), Freiburg: Herder, 1828. 2) *Darstellende Geometrie* (trans. G. Haussner), Leipzig: Engelsmann, 1900 (*Ostwalds Klassiker der exacten Wissenschaften*, no. 117).

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

Italian translation of the 4th ed. *Trattato della geometria descrittiva* (trans J. Corridi), Florence: Ricordi, 1838.

Russian translation. Nachertatel'nya geometriĭ (trans. V.F. Gazé, with the collaboration of T.P. Kraviets, D.I. Kargine and L.M. Loukowskaïa), Moscow: Academy, 1947.

Related articles: Poncelet (§27), von Staudt (§33), Klein (§42).

1 INTRODUCTION

When Gaspard Monge (1746–1818) gave his first set of lectures on descriptive geometry in Paris in 1795, no one other than himself had any idea what lay behind this title. In Year III of the revolutionary calendar, Monge succeeded in getting descriptive geometry introduced as a discipline that future teachers would have to study at the new *Ecole Normale*. He also made it the supreme discipline of what was to become the *Ecole Polytechnique* by allotting it half of the lecturing time [Paul, 1980, chs. 2–3]. Yet this discipline was not as new as it might have appeared. Coming out of the first lecture given by his colleague at the *Ecole Normale*, J.L. Lagrange exclaimed, 'I did not know I knew descriptive geometry' [de La Gournerie, 1855, 24].

The best way to find out what descriptive geometry is about is to 'listen' to Monge himself, whose words were carefully recorded by shorthand: 'The purpose of this art is two-fold. First it allows one to represent three-dimensional objects susceptible of being rigorously defined on a two-dimensional drawing. [. . .] Second [. . .] by taking the description of such objects to its logical conclusion, we can deduce something about their shape and relative positioning' (Programme of his lectures).

In the prologue to his twelve lectures, which were to be the starting-point of the interest of French mathematicians in geometry and of the upheaval mathematics underwent in the 19th century, Monge defined descriptive geometry as an 'art'. It is a 'science' replied in echo Michel Chasles (1793–1880) in his *Aperçu historique sur l'origine et le développement des méthodes en géométrie* before pursuing word for word with the rest of Monge's definition [Chasles, 1837, 189]. But at the same time, Chasles refused to admit that by itself descriptive geometry had the power to demonstrate fundamental geometrical properties such as whether a curve is planar or not.

This article is dedicated to this 'science' that can demonstrate nothing, or this 'art' that can be said to have provoked an upheaval in mathematics. Indeed, the two visions are not incompatible. As a geometrical method for depicting space, descriptive geometry can be seen both as a graphical technique and as a branch of geometry *per se*. But rather than attempting to place descriptive geometry between art and science, it is perhaps more profitable in the first instance to consider it as a language—which is also what Monge invites us to do. It is 'a language necessary for the engineer to conceive a project, for those who are to manage its execution, and finally for the artists who must create the different components' (Programme). It is, as it were, a language to speak 'space in three dimensions', at least when space is populated with objects 'that can be rigorously defined'.

2 GASPARD MONGE

2.1 Monge at Mézières

By pure chance in 1764, Monge entered through the back door of one of the most prestigious European engineering schools of the second half of the 18th century, the *Ecole du Génie* at Mézières. Just turned 18 years, his curriculum in a nutshell consisted in brilliant studies in Beaune, his native town, and then at Lyon. During the summer of 1764 he effected a survey of Beaune and drew a plan of it. The school's second in command, who happened to be visiting the town at the time, commended Monge for this work and recruited him to work at Mézières.

Little by little, Monge took over all the science teaching at the *Ecole du Génie*. Beginning as an assistant, he eventually replaced the mathematics professor, the Abbot Bossut, and from 1770 he took charge also of the physics lectures. In addition, he taught drawing, perspective and shadowing, as well as stone cutting. In 1775, he earned himself the title of 'Royal Professor of Mathematics and Physics'.

After having been elected as correspondent of Bossut at the Paris *Académie des Sciences* in 1772, Monge participated in several sessions of the *Académie* and came in contact with the Marquis de Condorcet, A.-A. Lavoisier and A.T. Vandermonde among others. Between 1771 and 1780, he presented eight memoirs, five of which were in analysis (essentially about partial differential equations), and three on differential geometry. Elected 'Associate Geometrician' of the *Académie* in 1780, he left the Mézières school in 1784 and settled in Paris. More interested at that time in physics and chemistry than in mathematics, he actively participated in the studies conducted by chemists in Lavoisier's immediate circle. Indeed, he succeeded in obtaining the synthesis of a small amount of water shortly after Lavoisier.

2.2 Monge's pedagogical projects

Monge committed himself body and soul to the revolutionary cause, and his political views were to become more radical in the course of the revolution. The Legislative Assembly elected him Navy Minister immediately after 10 August 1792 (which marks the fall of the monarchy), but he handed in his resignation eight months later. Nevertheless, he continued to participate actively in the revolutionary movement and the Public Welfare Committee's war effort. But his most important 'revolutionary' activity had to do with his participation in the pedagogical debates of the time and their consequences.

Monge was the main architect of the *Ecole Polytechnique*, and the creation of the school represents his most striking participation in the pedagogical projects of the Revolution. The school was destined to become the one training place for military and civil engineers and thus replaced in role the *Ecole du Génie* at Mézières and the *Ecole des Ponts et Chaussées* in Paris, both of which it emulated to a large extent. However, the number of students could not be compared with that of its predecessors under the *ancien régime*: nearly from 400 students were recruited in the first year. The Monge lectures we have been left with and which are discussed below are those that he gave at the *Ecole Normale*. But it is when looking at the way he organized his teaching at the *Ecole Polytechnique* that we can best

Figure 1. Monge, sketched by student L.M.J. Atthalin during a lecture at the *Ecole Polytechnique*, 1802 or 1803 (*Ecole Polytechnique* Archives; photograph by I. Grattan-Guinness).

assess his intentions concerning descriptive geometry. Figure 1 shows him drawn by a student there in about 1803.

Starting on 1 germinal (21 March 1795), Monge gave 34 lectures, that were abruptly interrupted on 7 prairial (26 May). On 8 thermidor (26 July), he resumed his lectures on descriptive geometry as applied to the cutting of wood and stone, perspective and shadowing, all at the fast rhythm of six sessions a 'decade' (the ten-day revolutionary week) until the beginning of year IV (the end of October 1795). After that date, he entrusted his colleague Jean Nicolas Pierre Hachette (1769–1834) with the full responsibility of teaching the descriptive geometry course.

As Monge became more and more involved with the politics of Napoleon, he proportionally disengaged himself from the school. From Napoleon's Italian Campaign in 1797 up until the time that he proclaimed himself Emperor in 1804, Monge was entrusted with a large number of official missions. He accompanied Napoleon in Egypt and became president of the Egyptian Institute. Elected Senator after the coup of 18 brumaire, Napoleon made him Senator of Liège in 1803, and he was to become President of the Senate from 1806 to 1807. But in the period of the Restoration (1815–1816) he was excluded from the *Académie* and died in July 1818.

3 THE SUBJECT MATTER OF MONGE'S LECTURES

The contents of Monge's lectures are summarised in Table 1. The lectures begin with a presentation of the conventions used to represent spatial bodies, then continue with a se-

Table 1. Summary by Lectures of Monge's book. The dates are given both in the Revolutionary and normal calendars.

Dates	**Lects.**	**Pp.**	**Topics**
1 pluviose – 1 ventose, Year III (20 January – 19 February 1795)	1–4, pt. 1; and debates	72	Programme. Preliminary considerations, problems about straight lines and planes.
1, 11 ventose (19 February, 1 March)	4, parts 2–5	46	Planes tangent to curved surfaces. Examples of the use of three-dimensional geometry to solve planar geometry problems.
21 ventose (11 March)	6	46	Intersection of curved surfaces and curves of double curvature.
1 germinal (21 March)	7	25	Application of surface intersection to the resolution of various problems: sphere inscribed in a pyramid, layout of a point from three sightings.
11, 21 germinal (31 March, 10 April)	8–9	36	Introduction to differential geometry.
1,11,21 floréal (20, 30 April, 10 May)	10–12		Theory of perspective and shadowing. [Published by Brisson in his 1820 edition, 138–187.]

ries of solved problems that are sometimes interlaced with more general considerations. The theoretical part of the course is subdivided into five chapters: 'Preliminary considerations', planes tangent to curved surfaces, curves of double curvature, 'The application of surface intersection to the resolution of various problems', and an introduction to differential geometry. In addition, he devoted three lectures to the theory of perspective and shadowing.

3.1 Preliminary considerations

The first part of the course is very revealing of Monge's general conception of descriptive geometry, for it starts with a lengthy lecture on the possible ways of characterizing a point in space *a priori*. He points out that only two planes are required in order to plot spatial objects as long as one introduces the notion of projection as opposed to that of distance as in analytical geometry. This is where he presents the basic principle of descriptive geometry: given two orthogonal planes, each point in space can be defined in terms of its projections onto these planes. When the two reference planes are folded on top of each other, one obtains on a flat sheet of paper what is known as the 'projected point diagram', that is to say, the two points in the plane that define the point in space (Figure 2).

The projection method indeed allows one to represent polyhedra, the projections of which are entirely determined by the projections of their vertices. But for non-specific surfaces, it is necessary to choose an extra convention and provide the method to construct

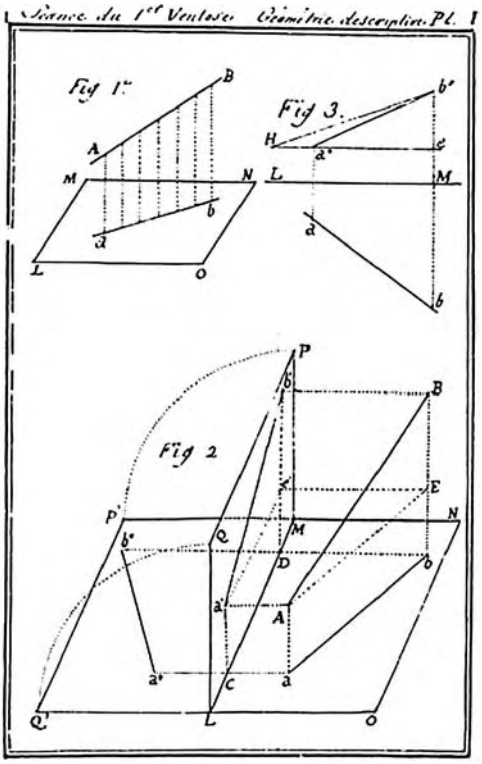


Figure 2. The principle of space representation in descriptive geometry: representation of a point and of a line segment.

the horizontal and vertical projections of two different generators that go through a single point in that surface. Monge gives a few examples of surfaces that can be defined in this way (cylinders, cones and revolution surfaces), and then treats the case of the plane in the same way. He defines the plane like any other surface, the only difference being that the generators that define it, straight lines, are simpler. This order of presentation was never used again in later works.

3.2 Tangential surfaces

The first part of the course ends with eight problems about straight lines and planes: tracing the line perpendicular to a given plane and passing through a given point, tracing a plane perpendicular to a given line and passing through a given point, and so on. The second part focuses on tangential planes and the perpendiculars to curved surfaces. Monge naturally begins with the simplest examples: constructing a tangential plane that goes through a single point of the surface of a cylinder, then a cone. Then he surprises his public and the reader by determining the distance between two lines and their common perpendicular. This question is certainly the most interesting problem of elementary descriptive geometry.

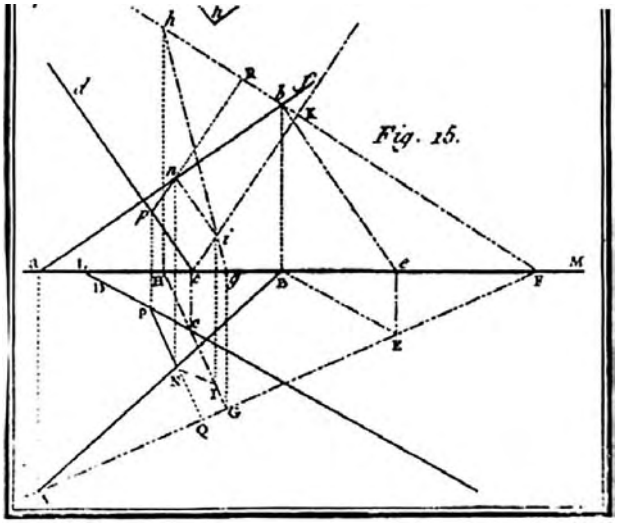


Figure 3. Distance between two straight lines. First, Monge defines the horizontal and frontal tracing of the plane containing the first given straight line (AB, ab) that is parallel to the second line (CD, cd). Then he constructs the contact directrix of the cylinder tangent to this plane, the axis of which is the line (CD, cd). To do this, he defines the projection (J, i) from any point (Point (C, c) on the diagram) of this axis onto the plane he has traced. The contact directrix is of course the line—parallel to the axis—that goes through point (J, i). It cuts line (AB, ab) at a point that belongs to the common perpendicular, which is therefore completely defined (PN, pn) since its direction is already known. The true magnitude of the distance between the two lines, which is not traced on the diagram, turns out to be the true length of segment $[PN, pn]$.

But while it can be resolved by considering only lines and planes and should therefore be approached in the preceding section, Monge treats it in terms of the definition of a revolution cylinder of given axis, tangent to a given plane (parallel to this axis). He presents it, therefore, as the reciprocal of the construction of the plane tangent to a cylinder (Figure 3).

Similarly, this surprising use of auxiliary surfaces allows Monge to deduce two planar geometry theorems from spatial geometry constructs. In the first instance, he demonstrates a theorem of Philippe de La Hire (Figure 4). The principle of Monge's demonstration consists in considering the planar geometry figure as the planar projection of three-dimensional space volumes. A circle is seen as the projection of a sphere, the two tangents to a circle as the generators of a cone. This demonstration, one of the most brilliant examples of the use of three-dimensional geometry to solve a planar geometry problem, brings one directly to the theory of poles and polars, which will be at the heart of work of J.V. Poncelet (1788–1867). Monge generalizes this theorem whilst considering any conical shape. Then, using the same method, he demonstrates the theorem proving (in modern terms) that the homothety centers which change two by two three circles are on a line.

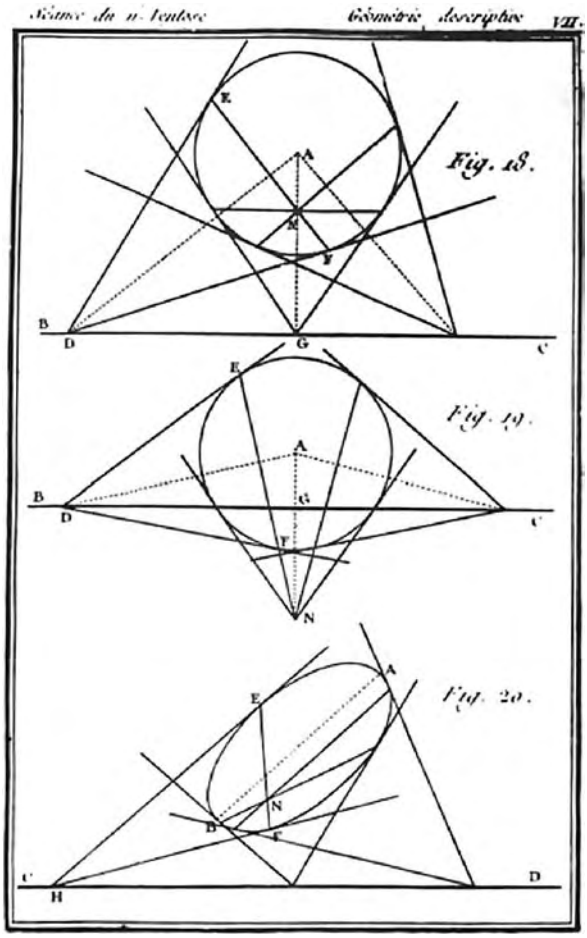


Figure 4. ‘Poles and polars’. The chord joining the points where tangents derived from a given point enter into contact with a circle pass through a fixed point when the point moves on a given straight line. Conversely, the tangents derived from the points of intersection of a straight line Δ and the circle cut one another at a point that moves along a straight line if Δ turns around a fixed point. Let Π be the plane defined by the straight line Δ and the centre of the circle A. Monge considers the sphere centred at point A with the same radius as the circle, and the cones of revolution tangent to the sphere whose vertex moves along the straight line Δ. The cones and the sphere admit the same tangent plane P, containing the straight line Δ (Π is a plane of symmetry of the figure and for the rest of the argument we may only consider the volumes situated ‘above’ Π). The point N where P comes into contact with the sphere belongs to all the circles of contact between the cones and the sphere; these circles are always situated on the planes perpendicular to Π. If these volumes are projected onto Π, the circles of contact are projected on the chords of the circle which pass through N projection of N, thus making it possible to deduce the theorem.

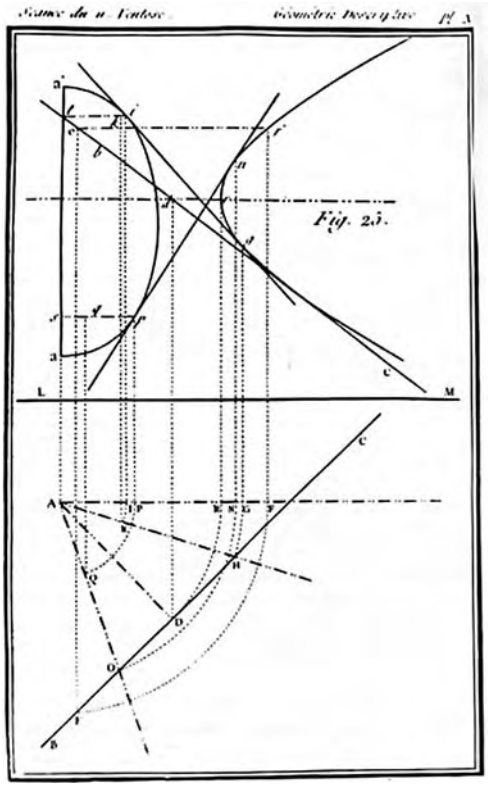


Figure 5. Plane tangent to a surface of revolution passing through a given straight line. Considering the tangent plane he is looking for, Monge supposes it to be rotating according to the motion that generates the surface of revolution. The straight line, included in the plane and labeled (BC, bc) in the Figure, will then generate a rotational hyperboloid. Monge first shows that the plane tangent to the first is also tangent to the second surface of revolution. He then determines point by point the intersection of the rotational hyperboloid with the frontal plane that contains the axis of the first surface of revolution. He finishes off the construction using the tangents common to the hyperbole so defined and the directrix of the first surface of revolution.

Monge ends this section with a far more delicate problem: constructing the plane tangent to a revolution surface passing through a given straight line (Figure 5). This example, like the one concerning the distance between two straight lines, is very revealing about Monge's teaching. It is mainly for him an opportunity to display the gamut of possible auxiliary surfaces, to show that they are not limited to planes, cones and cylinders.

3.3 Curves of double curvature and differential geometry

The third part of the course focuses on the intersection of curved surfaces and double curvature curves. Monge takes this opportunity to present the method known as the 'auxiliary planes' method. This consists in having a set of planes intervene, the intersection of these planes with each surface being geometrically defined so that each of the auxiliary planes allows one to construct one (and possibly more) points along the curve of intersection (see Figure 6).

Monge then gives several applications of surface intersection. The two following Lectures, which form the fifth part of the course, do not concern descriptive geometry according to today's nomenclature but some of the results that he had published in some of his memoirs for the *Académie*. In the first of these lectures, appealing to visual and intuitive comprehension, he presents the notion of the evolute of a planar curve as the generalization

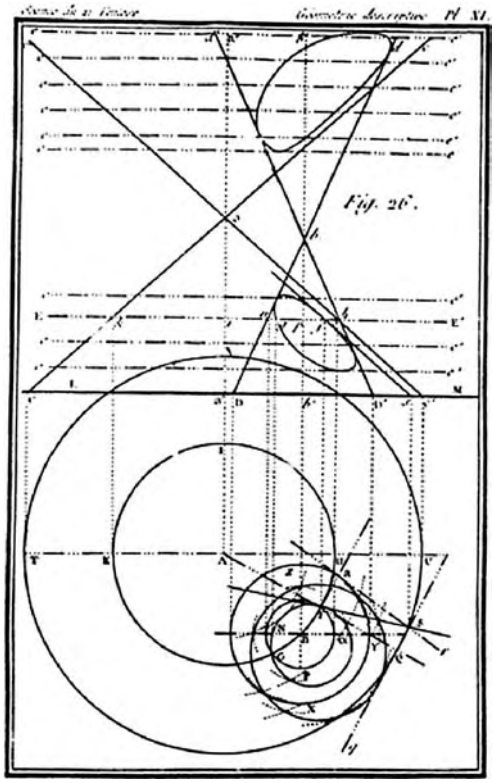


Figure 6. Intersection of two cones; auxiliary plane method.

of a circle, the involute playing the role of the centre (Figure 7). Conversely, the construction of the involute starting with the evolute of a planar curve allows him to introduce the notions of radius of curvature and center of curvature. He defines the polar surface as being the envelope surface of planes normal to the curve (Figure 7). At this point, Monge introduces the notion of developable surface and cuspidal edge. He ends this Lecture by showing that the perpendiculars to a given surface along a curvature line generate a surface that can be developed.

3.4 Shadowing and perspective

The theoretical section ends with this complement of differential geometry. However, it does not end the course on descriptive geometry as a whole.

At the *Ecole Normale*, Monge ended up only giving the lectures on shadowing and perspective. But at the *Ecole Polytechnique*, he also taught the applications of descriptive geometry to stone cutting and carpentry, the drawing up of plans and maps, and the technical drawing of machines and for architecture. At the time, descriptive geometry was thus

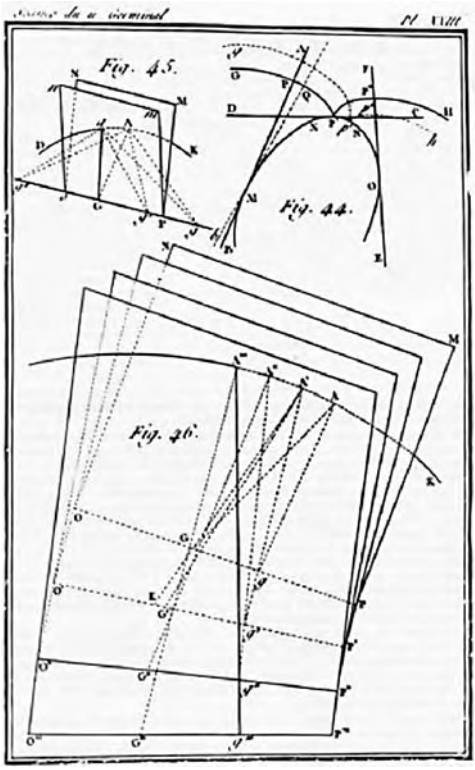


Figure 7. Evolute and involute of a curve; radius of curvature of a gauche curve; its polar surface.

defined in a much broader way than it is today, covering a very large number of subjects (see Figures 8 and 9 as examples).

4 THE PRINCIPAL AIMS OF MONGE'S COURSE

4.1 Descriptive and practical geometry

Monge never presented descriptive geometry as a new science of which he might be the founding father. Quite the contrary, he describes it as 'having been practiced for a great deal longer [than Analysis] and by many more people'. He even adds that descriptive geometry having been practiced 'by men whose time was precious, the (graphical) procedures were simplified and, instead of considering three planes, one got—thanks to projections—to only require two planes explicitly' (p. 312). Thus, contrary to what is later going to be said against Monge: the minimalist character of the diagram lines used in descriptive geometry is not the fruit of a mathematician's theoretical research but stems from the perfecting of practices over the years. Although he does not cite any names, he is obviously referring to

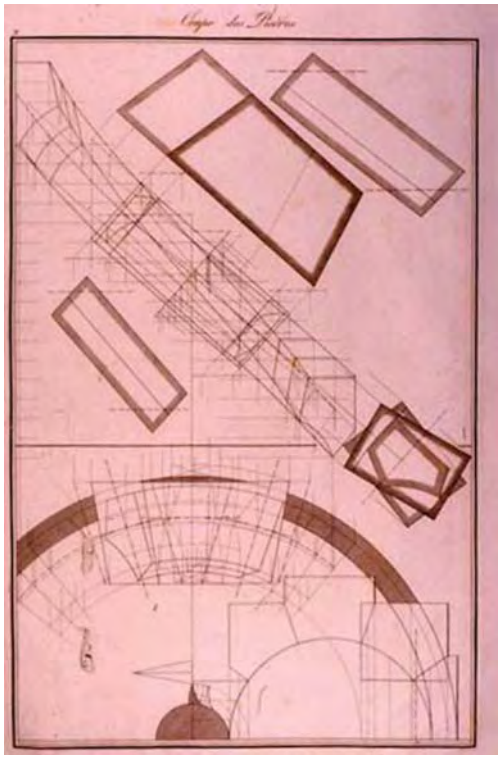


Figure 8. The see-through Saint-Gilles 'Vis', or spiral staircase. Etched diagram from the *Ecole Polytechnique* Archives, files for year III.

the drawings of stone-cutters and carpenters. The privileged ties that descriptive geometry enjoys with various graphical techniques is made evident by the abundant examples that he gives in the foundation course, which is constantly enriched by references to diverse techniques that are likely to use descriptive geometry.

4.2 Descriptive geometry and analysis

Monge also returns on several occasions in his lectures to the analogies that exist between descriptive geometry and analysis. He already touches upon this theme in the second Lecture: 'it is not without reason that we are comparing here descriptive geometry and algebra; the two sciences are very closely related. There is no descriptive geometric construct that cannot be transposed in terms of Analysis; and when the problem does not involve more than three unknowns, each analytical operation can be regarded as the script of a play on the geometrical stage' (p. 317).

Monge draws the logical conclusion from this analogy and focuses on it on several occasions: 'it is desirable that the two sciences be cultivated together: descriptive geometry can bring to the most complicated analytical operations the obviousness that characterizes

Figure 9. Architectural drawing. From the drawings portfolio of student J.-B.-C. Dalesme (1812 promotion), *Ecole Polytechnique* Archives.

it and, in turn, analysis can bring to geometry the trait of generality which is its essence' (p. 317). In this parallel Monge's philosophy is best expressed. He indeed tried to put it into practice at the *Ecole Polytechnique* where he simultaneously taught descriptive geometry and analysis as applied to geometry (the latter in [Monge, 1795]).

4.3 'Properties of surfaces'

'My aim [. . .] is to get you acquainted with the properties of surfaces', declares Monge on one occasion in a discussion with his students (p. 321). This sentence is probably the best description of this set of lectures. Several elements are brought together to achieve this aim.

First it must be noted that the space imparted in Monge's lectures to the problems of lines and planes is very limited. Descriptive geometry begins with the manipulation of surfaces; it is a tool that allows them to be introduced, conceived, used in proofs and represented.

Faced with the number of solutions that offer themselves, Monge always chooses the most graphic. Cylinders, cones, spheres or other hyperboloids fill the space, providing

matter for the speaker to work from, a support for the listener or the reader's intuition and a certain substance for the demonstrations which, without them, would have been less captivating. The subtle play of auxiliary surfaces, which he manipulates with a consummate art, allows him to turn the problems around and systematically study each problem and its reciprocal. Determining the distance between two straight lines as the reciprocal of the problem of determining the plane tangent to a cylinder is one example. But the most superb illustration of turning the situation around, and the richest from the geometrical point of view, is given in the demonstrations in planar geometry that use descriptive geometry.

Another characteristic element of this course is the way in which Monge expresses the relationship between geometrical reasoning and its graphical translation, its 'representation'. For example, in determining the plane tangent to a surface of revolution (Figure 5), neither the surface nor the tangent plane or the rotating hyperboloid appears explicitly. Similarly, in determining the distance between two straight lines, the cylinder, which is present in the demonstration, is totally absent from the projection diagram (Figure 3); the only element of the surface that has been kept is the one that effectively plays a role, and that is the contact line.

Being pared down as much as possible, the drawing does not show the objects but merely the geometrical constructs used in the reasoning, constructs that would have been drowned and indecipherable had the various surfaces been represented. The projection diagram in descriptive geometry forces one to choose the elements that are needed for the geometric proof. 'The old geometry bristles with diagrams. The reason is simple. Because there was a lack of general abstract principles, each problem could only be analyzed from a concrete standpoint, using the very figure that was the object of the problem. It was only by looking at this figure that one might discover the elements necessary for the proof or the solution one sought', wrotes Chasles. He even adds, much to the reader's surprise, 'no one has surpassed Monge in conceiving and doing geometry without using figures' [Chasles, 1837, 208]. He points here to one of the riches of Monge's course and highlights the paradoxical contribution of descriptive geometry.

By the very content of his lectures, Monge therefore goes far beyond the narrow and relatively restricting framework he had given himself when, in his introduction, he defined descriptive geometry as a graphical technique. 'And if there is someone amongst you whose [. . .] heart begins to beat, that is it, he is a geometrician', he declared during one of the debate sessions (p. 321). There is no doubt that his lectures made the heart of many a student beat, and thereby he transformed a whole generation of *Ecole Polytechnique* students into geometers.

5 THE INFLUENCE OF MONGE'S LECTURES

5.1 The reputation of descriptive geometry

The teaching of descriptive geometry developed rapidly. In France, Hachette was to be the most ardent promoter of the Monge theory, which he taught not only at the *Ecole Polytechnique* but also at the Paris *Faculté des Sciences* and at the *Ecole Normale* from 1810 onwards. He also produced new editions of Monge's lectures, a work that was translated into several languages, as the publication history above shows.

In giving a panorama of the history of geometrical methods from antiquity to his time, J.-L. Coolidge treats descriptive geometry with circumspection. While recognizing its technical role, he reduces its scientific value to something of little significance: 'It is hard to point to important properties of space figures which were first found by the methods of Monge or which are more easily proved by those methods than by others' [Coolidge, 1940, 112–113]. This judgement seems rather excessive even if, in the hopes that Monge had placed in the new discipline that he had created, there was something of a revolutionary utopia that was soon to disappear.

Certainly, the theorems on the joining of gauche surfaces or on determining the full shadow separator of the triangular thread screw have neither revolutionized mathematics nor bowled over mathematicians. Nevertheless, Monge's lectures played an important part in the change in mentality that took place at the beginning of the 19th century among mathematicians. They became aware that 'Geometry, which had been looked upon for a century as powerless by itself and having to draw all its resources and acquisitions from algebra, could on the contrary be a source of general principles and methods as fertile as those of algebra, that these methods sometimes had certain advantages in allowing one to penetrate all the way to the origin of truths and lay bare the mysterious chain that links them to each other' [Chasles, 1870, 81].

Three essential ideas appear in Monge's lectures and will be developed afterwards: the notion of projection and transformation, the modification of the relationship between algebra and geometry, and the implicit use of what Poncelet was to call 'the principle of continuity' (§27.1.2). Let us briefly consider them.

5.2 Projections and transformations

'When thinking carefully about the main advantage of descriptive geometry and the coordinate method, and reflecting upon why these branches of mathematics seem to be akin to absolute doctrines, the principles of which are few but related and linked in a necessary manner and uniform progression, it is not long before one realizes that this is solely due to the use they make of projection' [Poncelet, 1822, 28].

At the heart of descriptive geometry is of course the use of the notion of projection in order to represent points and surfaces from space. But descriptive geometry also allows one to make 'the intimate and systematic link between three-dimensional and planar figures' [Chasles, 1837, 191]. It is in the handling of reciprocal relationships that the true riches of the notions of projection and geometrical transformation become really apparent. C.J. Brianchon, followed by Poncelet, would later successfully cultivate this method, which is one of the hallmarks of the 'Monge School'.

5.3 Geometrical intuition

The concern and the desire to regain from algebra the terrain that had hitherto escaped geometry are constantly present in the various descriptive geometry treatises and theses of Monge's successors. Felix Klein, who declares 'having been raised [...], thanks to [his] professor, Plücker, in the Monge tradition', considers that one of the major contributions of this tradition was 'the application of geometrical intuition to algebra' (quoted in

[Taton, 1951, 240]). He even adds in *The Erlangen Programme* [Klein, 1872]: 'One must not do away with the prescription that a mathematical problem should not be considered to have been exhaustively examined as long as it has not become intuitively obvious. To discover something by means of algebra is indeed a very important step, but it is only the first step' (compare §42).

5.4 The principle of continuity

The 'intimate fusion' [Poncelet, 1822, xx] of two ways of proving a particular property allows one to bring geometrical intuition to the analytical method. But the fact that analytical demonstrations, established in the case of real elements, extend to cases in which some of the elements become imaginary, directly leads Monge to admit that associated geometrical demonstrations must also be extended under the same conditions. In the theorem about poles and polars (Figure 3), there are two distinct cases to be considered *a priori*. The plane tangent to the sphere and including the given straight line only really exists if the line does not intersect with the given circle. Monge indeed traces both figures but makes no distinction in the corpus of the demonstration, apparently taking the fact that the tangent plane might be real or imaginary as negligible and using for the first time the principle of continuity.

6 CONCLUSION

Created to 'pull the French nation out of its hitherto dependence on foreign industry' (Programme, 305), descriptive geometry will paradoxically have had more influence in the field of mathematics than in the technical world—contrary to Coolidge's assertion.

Descriptive geometry has been two-faceted from the time it was created. It is on the one hand an entirely new discipline, a 'revolutionary' discipline that acquires a name, and sees its object and place in mathematics defined in Monge's lectures. It offers an unprecedented manner of tackling three-dimensional geometry or, to be more exact, linking planar geometry with spatial geometry. It is also revolutionary because of the position it can aspire to in the school system, in the training of the elite as in general technical training. But it simultaneously appears as the last stage of a tradition that is losing momentum, as the ultimate perfecting of previous graphical techniques and in that capacity, marks the endpoint of an evolutionary process as much as the birth of a new branch of geometry. As such, it can also be viewed as a transition discipline that allowed a gentle evolution to take place: from the 'artist engineer' of the Old Regime, whose training was based on the art of drawing rather than scientific learning, to the 'learned engineer' of the 19th century for whom mathematics—and algebra in particular—is going to become the main pillar of his training.

BIBLIOGRAPHY

Belhoste, B., Dahan-Dalmedico, A. and Picon, A. (eds.) 1994. *La formation polytechnicienne, deux siècles d'histoire*, Paris: Dunod.

Booker, P.J. 1963. *A history of engineering drawing*, London: Northgate. [Repr. 1979.]

Coolidge, J.L. 1940. *A history of geometrical methods*, Oxford: Clarendon Press. [Repr. New York: Dover, 1963.]

Chasles, M. 1837. *Aperçu historique sur l'origine et le développement des méthodes en géométrie*, Bruxelles: Hayez. [Repr. Paris: Gabay, 1989.]

Chasles, M. 1870. *Rapport sur les progrès de la géométrie*, Paris: Gauthier–Villars.

de La Gournerie, J. 1855. *Discours sur l'art du trait et la géométrie descriptive*, Paris: Mallet-Bachelier.

Klein, F. 1872. *Vergleichende Betrachtungen über neuere geometrische Forschungen*. Erlangen: Deichert. [Repr. in *Mathematische Annalen, 43* (1893), 63-100. See §42.]

Loria, G. 1908. 'Perspektive und darstellende Geometrie', in M. Cantor (ed.), *Vorlesungen über Geschichte der Mathematik*, vol. 4, Berlin: Teubner, 577–637.

Loria, G. 1921. *Storia della geometria descrittiva delle origini, sino ai giorni nostri*, Milan: Hoepli.

Monge, G. 1795. *Feuilles d'analyse appliquée à la géométrie à l'usage de l'Ecole Polytechnique*, Paris: *l'Ecole Polytechnique*. [Later eds. from 1807 under the title *Applications de l'analyse à la géométrie*, Paris: Bernard.]

Paul, M. 1980. *Gaspard Monges "Géométrie descriptive" und die* Ecole Polytechnique, Bielefeld: University.

Poncelet, J.-V. 1822. *Traité des propriétés projectives des figures*, Paris: Bachelier. [See §27.]

Sakarovitch, J. 1998. *Epures d'architecture, de la coupe des pierres à la géométrie descriptive, XVI^{e} – XIX^{e} siècles*, Basel: Birkhäuser.

Taton, R. 1951. *L'œuvre scientifique de Monge*, Paris: P.U.F.

Taton, R. (ed.) 1964. *Enseignement et diffusion des sciences en France au XVIIIe siècle*, Paris: Hermann. [Repr. 1986.]

CHAPTER 18

P.S. LAPLACE, *EXPOSITION DU SYSTÈME DU MONDE*, FIRST EDITION (1796); *TRAITÉ DE MÉCANIQUE CÉLESTE* (1799–1823/1827)

I. Grattan-Guinness

The *Traité* was an authoritative statement on celestial and planetary mechanics of its time, and also an important source on several new mathematical methods. The *Exposition* gave a non-technical account of the mechanics and related physics, and gained a wide readership over several editions.

The changes in publisher below indicate commercial successorship; all volumes were published in Paris. Revolutionary year dates are converted to the normal system.

Exposition du système du monde

First publication. 1st ed. 2 vols., (printed) Cercle-Social, 1796. 314 + 312 pages.

Later editions. 2nd ed. 1799, Duprat. 3rd ed. Courcier, 1808. 4th ed. Courcier, 1814, 1500 copies. 5th ed. Bachelier, 1824. [Repr. as *Oeuvres*, vol. 6, Imprimerie Royale, 1846.]

Sixth edition. Bachelier, 1835, 1000 copies in octavo and quarto formats each. [Posthumous, using the organisational changes envisaged by Laplace. Repr. as *Oeuvres complètes*, vol. 6, Gauthiers–Villars, 1884 (photorepr. Hildesheim: Olms, 1966). Octavo photorepr. Fayard, 1984.]

Manuscripts. Of 3rd ed., *Bibliothèque de l'Observatoire*, Paris.

German translation. Of 1st ed., *Darstellung des Weltsystems* (trans. J.K.F. Hauff), 2 vols., Frankfurt am Main: Varrentrapp und Wenner, 1797.

English translations. 1) Of 3rd ed., *The system of the world* (trans. J. Pond), 2 vols., London: Phillips, 1809. 2) Of 5th ed., *The system of the world* (trans. H. Harte), 2 vols., London: Longmans, 1830.

Spanish translation. Of 6th ed., *Breve historia de la astronomia* (trans. J. Panfu and A.B. Besco), Buenos Aires: Espasa-Calpe, 1947.

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

Traité de mécanique céleste

Volume	Year	Month	Publisher	Pages	Print-run
1	1799	September?	Duprat	xxxi + 368	1500
2	1799	September?	Duprat	382	1500
3	1802	December?	Duprat	xxvi + 303	
Supplement	1808	September?	Courcier	24	
4	1805	May?	Courcier	xl + 347	
Supplement 1	1806	June?	Courcier	63	
Supplement 2	1807	July?	Courcier	78	
5	1823–1825	–	Bachelier	ix + 420	1000
Supplement	1827	?	Bachelier	30	

Reprints. Vols. 1–4 and the three supplements: Bachelier, 1829–1839. 350 copies. All vols. and supplements: as *Oeuvres*, vols. 1–5, Imprimerie Royale, 1843–1846. Also as *Oeuvres complètes*, vols. 1–5, Gauthiers–Villars, 1878–1882. [Latter ed. photorepr. Hildesheim: Olms, 1966; vol. 5 repr. New York: Chelsea, 1966, with Bowditch below.]

Photoreprint. Brussels: Culture et Civilisation, 1967.

Manuscripts. Supplement to vol. 5: *Bibliothèque de l'Observatoire*, Paris, ms. 1520.

German translation. Of vols. 1–2: *Mechanik des Himmels* (trans. J.C. Burckhardt), 2 vols., Berlin: La Garde, 1800–1802.

English translations. 1) Of Book 1: *A treatise upon analytical mechanics* (trans. J. Toplis), Nottingham (printed), London: Longmans etc., 1814. 2) Of vol. 1: *A treatise of celestial mechanics* (trans. H. Harte), 2 vols., Dublin: Milliken, 1822–1827. [Vol. 1 also London: Longmans etc.] 3) Of vols. 1–4: *Mécanique céleste by P.S. Laplace* (trans. N. Bowditch), 4 vols., Boston: Hillard etc., 1829–1839. [Photorepr. as *Celestial mechanics*, New York: Chelsea, 1966 (with obituary moved from vol. 4 to vol. 1).]

Related articles. Newton (§5), d'Alembert (§11), Lagrange (§16, §19), Gauss on astronomy (§23), Laplace on probability (§24), Lacroix (§20), Green (§30), Cauchy on real-variable analysis (§25).

1 BACKGROUND

Born into the bourgeoisie in Normandy, Pierre Simon Laplace (1749–1827) exhibited his mathematical powers early; in his late teens he went to Paris, where he was based for the rest of his life [Gillispie, 1998]. He gained the attention of Jean d'Alembert and the Marquis de Condorcet, and began to work in the calculus and differential equations, difference equations, and aspects of mathematical probability. Soon after gaining election to the *Académie des Sciences* in 1773, he became seriously interested in mathematical astronomy, which was to become his dominant concern. By then two major areas of activity were evident: in celestial mechanics, the fine details of the motions of the heavenly bodies

as analysed using especially Newton's laws and allowing for perturbations; and in planetary mechanics the analysis of their shapes, especially that of the Earth following the demonstration of its oblateness in the 1740s, and consequent topics such as the motion of the sea and of tides, and the analysis of projectiles. The main tool was the calculus, including series, functions, and difference and differential equations, which themselves were importantly advanced.

These areas were the chief ones in mathematics of the time; in tackling them Laplace conjoined with the leaders of the mathematical world, not only his mentors but also Leonhard Euler (1707–1783) and J.L. Lagrange (1736–1813), the latter then based in Berlin but to move to Paris in 1787. His relationship with Lagrange seems to have been correct though rather formal. Unease is evident in his relationship with his younger kinsman Adrian-Marie Legendre (1752–1833), especially in planetary mechanics and related mathematical methods.

Between 1770 and 1800 a mass of work was produced in mathematical astronomy and the calculus by these three men, and some other figures. Lagrange's main product was his *Méchanique analitique* (§16), published in 1788 in Paris though written during his time in Berlin. Laplace's contributions mostly appeared as papers, and also in his first book, *Théorie du mouvement et de la figure elliptique des planètes* (1784), whose obscurity owes much to the failure of the editors of his "collected" works to reprint it. The history of these developments is very complicated and intertwined: many details can be retrieved from the Bowditch translation listed above. See also [Gautier, 1817; Wilson, 1980, 1985; Grattan-Guinness, 1990, chs. 4–6; Grant, 1852], and [Taton and Wilson, 1995] for the general astronomical background.

2 THE *EXPOSITION*

A major change in everybody's life was the French Revolution of 1789. Professionally Laplace gained status when the *Bureau des Longitudes* was formed in 1795 as the national organisation to assist practical astronomy and navigation; for he was unofficially its leader until his death. In addition, when the *Ecole Polytechnique* was founded in 1794 as an elite school for training civil and military engineers Laplace was one of the founding graduation examiners (with Lagrange as a founding professor of 'analysis'); and in 1799, during his brief period under Bonaparte as Minister of the Interior he both urged the formation of a governing body for the school and became a very influential member of it.

During those years the *Exposition* was published and the *Traité de mécanique céleste* started to appear. In his late forties, Laplace seems to have felt ready to emulate Lagrange in writing an authoritative account, in his case of mathematical astronomy together with some new methods in the calculus. The history of the writing of both works is obscure, but apparently Laplace avoided the chaos following the Revolution by "disappearing" from public life: he sent a copy of the *Exposition* to the Swiss physicist George Lesage in 1797, and with striking frankness described it as 'the fruit of my retreat into the country, for the duration of this unhappy revolutionary government which has cost such tears for the true friends of France and of humanity' [Grattan-Guinness, 1990, 1281].

The *Exposition* had appeared in two volumes in 1796 (Table 1); I cite it by Book and chapter number as 'N#M'. The first two Books were descriptive of astronomical phenom-

Table 1. Contents of *Exposition*, 1st edition (1796)

Book; chs.	**Pages**	**'Title'**
1; 14	vol. 1, 9–168	'On the apparent motions of celestial bodies'.
2; 7	vol. 1, 169–237	'On the real motions of celestial bodies'.
3; 5	vol. 1, 238–314	'On the laws of motion'.
4; 15	vol. 2, 5–198	'On the theory of universal gravity'.
5; 6	vol. 2, 199–312	'Precis of the history of astronomy'.

ena, indicating cause in at most a general way. Then in Book 3 Laplace reviewed the main principles of mechanics that he wished to deploy. In Book 4 he worked through much of the material again in a somewhat more technical manner (for example, tides in 1#13 and 4#10). However, no formulas were given—and also no diagrams, surely another sign of influence from Lagrange, who felt that their use compromised the rigour and generality of theories (§16.2). There were also no references, though scientists were named quite frequently. Data were often given: Laplace used the centesimal division of the quadrant into degrees and decimal division of the day (10 hours, 10 minutes, 100 seconds), but rendered distances in terms of the toise.

Laplace wished to derive celestial mechanics from the law of universal gravitation with Newton's inverse-square law of attraction. However, Newton's laws and notions were not given the prominence that might be expected: instead Laplace exhibited the frequent Continental preferences, especially in Lagrange's mechanics, for d'Alembert's principle and those of least action and virtual velocities (3#2). In a striking passage he chose 'force' to name (mass $\times$ velocity), distinguished from 'accelerative force' and 'motive force'.

Most of Book 5 was devoted to the history of astronomy, from ancient to recent times. As history it is derivative—Laplace might have been advised by his colleague J.B.J. Delambre (1749–1822)—but its existence at all is very striking. In the final chapter he put forward a suggestion which has unfortunately eclipsed much of his other work despite his clear warning of its speculative character: the 'nebular hypothesis', that the Sun had been enveloped in a vast hot atmosphere that had condensed and cooled down to form fluid annuli, which in turn condensed into planets; satellites were formed similarly from the planets' own atmospheres. Comets were also so formed, but the large eccentricities of their orbits led most of them to extinction or absorption. Finally, distant light came from stars, clumped together in milliards in various parts of the Universe (5#6). He was an enthusiast for atmospheres, claiming that all planets possessed them although those with weaker gravities may lose them (4#9). Further considerations about the immensity of the Universe included a striking sentence that even hinted towards black holes (vol. 2, 305).

As the publication history show, the book was soon rendered into German, and there were five later French editions. Laplace changed or updated the text in places [Eisenstaedt, 1991]; for example, the nebular hypothesis ended up as a final appendixed note, and free from black holes. Most changes occurred in Book 4; the most substantial one is noted in section 9 below. But the design in five Books remained, often with little change of text. Somewhat eccentrically he also maintained the decimal measures, although the toise gave way to the metre after its introduction in 1799.

The existence of the later editions reflects its success; it is nicely written, although the general readers for whom it seems to have been intended would well have been tested. But the elite received a short summary, prepared by the engineer (and professor at the *Ecole Polytechnique*) G. Riche de Prony (1755–1843) for the *Institut de France* [de Prony, 1801], and two years later a course in astronomy started at that school with the *Exposition* as the set book.

3 THE 'FIRST PART' OF *MÉCANIQUE CÉLESTE*, 1799

At one point in the *Exposition* Laplace mentioned that the *Mécanique céleste* was written and would be published (vol. 2, 7). It appeared in 1799 in two volumes, constituting its 'First Part' and beginning with a very detailed table of their contents. The importance of the publication was expressed by publisher Duprat in a most singular way: all the sheets of paper used carried 'MECANIQUE CELESTE' as their watermark on the bottom. (The 1840s 'national' edition of his works was to offer a similar homage, with 'OEUVRES DE LAPLACE' and 'LOI DU 15 JUIN 1842' on alternating sheets.) As with the *Exposition* (and also his first book of 1784), the volumes were rendered into German in the early 1800s; and a French summary was produced, this time by his follower J.B. Biot (1774–1862) [Biot, 1802]. Two further volumes appeared in 1802 and 1805, not so honourably watermarked.

As with the *Exposition*, decimal measures were used (but now including the metre); with an exception to be noted in section 7, no diagrams were furnished; and many data were given, often calculated by Aléxis Bouvard, Laplace's assistant at the *Bureau*. By contrast, mathematics was everywhere, often fearfully: Bowditch memorably remarked that 'Whenever I meet in La Place with the words "Thus it plainly appears", I am sure that hours, and perhaps days, of hard study will alone enable me to discover *how* it plainly appears' (vol. 4, 62 of obituary). Unlike the earlier work, very few references were provided; Isaac Todhunter wittily surmised that Laplace 'supposed the erudition of his contemporaries would be sufficient to prevent them from ascribing to himself more than was justly due' [Todhunter, 1861, x–xi]. I cite the work by Book and article number, again in the form 'N#M'.

Table 2 shows the great ambition of the work: for example, the perturbations of the satellites of Jupiter (8#11–15), and attention in various places to Uranus, which had been identified only in 1781. However, the column of page numbers indicates that several chapters were very short. Only a severe selection of features and points can be made here; several more appear in [Todhunter, 1873, chs. 28–34].

4 *MÉCANIQUE CÉLESTE*, THE CELESTIAL VOLUME 1

The opening is odd in one respect. Laplace spoke of 'force' disturbing bodies from equilibrium, and seemed to mean (mass $\times$ acceleration); but (mass $\times$ velocity) was introduced in 1#5, and the various relationships rehearsed again in 1#24.

An important early theorem was silently obtained from a posthumous paper of 1793 by Euler on the linear combination of torques. Laplace used it and the theorem of conservation of areas to claim the existence of the 'invariable' plane of the planetary system, defined by

Table 2. Contents of *Mécanique céleste*, volumes 1–4 (1799–1805). The original first page numbers of chapters are given; quotations come from Bowditch.

Chs., arts.	**Page**	**'Titles' and topics**
Part 1, Volume 1		'General theories of the motions and figures of the heavenly bodies'.
Book 1		'On the general laws of equilibrium and motion'.
1, 1–3	3	Equilibrium and composition for forces on a point.
2, 4–12	14	'On the motion of a material point'.
3, 13–16	36	'On the equilibrium of a system of bodies'.
4, 17	47	'On the equilibrium of fluids'.
5, 18–23	50	'General principles of the motion of a system of bodies'.
6, 24	65	'Motion of a system of bodies'; force and velocity.
7, 25–31	70	Motions of any solid body; rotation, axes.
8, 32–37	91	'On the motion of fluids'; incompressibility; earth's atmosphere.
Book 2		Law of gravitation; motions of 'centres of gravity of the heavenly bodies'.
1, 1–6	111	'The law of universal gravity deduced from observation'.
2, 7–15	124	Differential equations for motion of a system of bodies; sphere, cylinder.
3, 16–25	154	'First approximation' to these motion; 'elliptical motion'.
4, 25–39	190	'Determination of the elements of the elliptical motion'.
5, 40–45	235	Methods of finding motions, 'by successive approximations'.
6, 46–52	254	'Second approximation' to motions; 'perturbations'.
7, 53–62	286	'Secular inequalities' of the motions; eccentricities and perihelia.
8, 63–73	321	'Second method of approximation' to motions. [End 368.]
Volume 2, Book 3		'On the figures of the heavenly bodies'.
1, 1–7	3	'Attractions of homogeneous spheroids' with second-order surfaces.
2, 8–17	23	'Development' of attraction in an infinite series; harmonic analysis.
3, 18–21	50	'Figure' of rotating homogeneous fluid in equilibrium.
4, 22–37	63	Figure of nearly spherical spheroid covered by fluid in equilibrium.
5, 38–43	109	Comparison of theory with observations.
6, 44–46	155	'On the figure of the ring of Saturn'.
7, 47	167	'On the figure of the atmospheres of the heavenly bodies'.
Book 4		'On the oscillations of the sea and atmosphere'.
1, 1–12	171	'Theory of the ebb and flow of the sea'.
2, 13–14	204	'Stability of the equilibrium of the sea'; density relations.
3, 15–20	212	'Ebb and flow' of the tides in ports.
4, 21–43	233	Comparison of theory with observations.
5, 44	294	'On the oscillations of the atmosphere'.
Book 5		'Motions of the heavenly bodies about their own centres of gravity'.
1, 1–14	299	Earth; differential equations, equinoxes; effect of the sea, inclination of axis.
2, 15–19	356	Moon; libration; 'motions of the nodes'.
3, 20–22	373	Rings of Saturn; differential equations; also Uranus. [End 382.]

Table 2. (*Continued*)

Chs., arts.	Page	'Titles' and topics
Part 2, Volume 3		'Particular theories of the motions of the heavenly bodies'.
Book 6		'Theory of the planetary motions'; short preface.
1, 1–11	5	Inequalities depending on powers of eccentricities and inclinations > 2.
2, 12–18	33	'Inequalities depending on the square of the disturbing force'.
3, 18′	55	'Perturbations depending on the ellipticity of the Sun'.
4, 19	58	Perturbations of planets because of 'the action of their satellites'.
5, 20	60	'Elliptical part of the radius vector'; mean motion of a planet.
6, 21–23	61	Numerical values of quantities in planetary inequalities.
7, 24–26	86	Numerical values of secular variations of planetary orbits.
8–9, 27–28	95	'Theory of' Mercury, Venus.
10, 29–31	103	'Theory of the Earth's motion'.
11–14, 32–38	115	'Theory of' Mars, Jupiter, Saturn, Uranus.
15, 39–42	147	'Equations of condition' for inequalities to verify numerical values.
16, 44	156	'On the masses of the planets and Moon'.
17, 45–46	162	Astronomical tables, and the invariable plane of our system.
18, 47	164	Effects of the fixed stars upon our system.
Book 7	169	'Theory of the Moon'; short preface.
1, 1–19	181	Integration of the differential equation of its motion; numerical values.
2, 20–21	250	Lunar inequalities due to oblateness of Earth and Moon.
3, 22	263	Lunar inequalities due action of the planets.
4, 23–25	273	Comparison of theory with observations.
5, 27–28	289	Apparent long-period lunar inequality.
6, 29–30	296	Secular variations of Moon and Earth maybe due to the Sun. [End 303.]
Supp., 1–5	24 pp.	On Poisson's analysis of second-order perturbations.
Volume 4		General preface, especially on motions of the satellites.
Book 8		'Theory of the satellites of Jupiter, Saturn and Uranus'.
1, 1–2	2	Satellites of Jupiter.
2, 3–5	8	Inequalities of these satellites depending upon eccentricities and inclinations.
3, 6–8	20	Inequalities of satellites depending upon eccentricities of orbits.
4, 9–11	32	Inequalities of satellites in latitude.
5, 12–13	50	Inequalities depending upon squares and products of eccentricities and inclinations.
6, 14–19	59	Inequalities depending upon the square of disturbing force.
7, 20–25	83	'Numerical values of the preceding inequalities'.
8, 26	105	'On the duration of the eclipse of any satellite'.
9, 27	121	'Masses of the satellites, and oblateness of Jupiter'.
10, 28	127	Eccentricities and inclinations of orbits of satellites.

Table 2. (*Continued*)

Chs., arts.	**Page**	**'Titles' and topics**
11–15, 29–33	135	Motions of the satellites of Jupiter.
16, 34	169	'On the duration of the eclipses of the satellites'.
17, 35–37	173	'On the satellites of Saturn'; positions in plane of ring.
18, 38	190	'On the satellites of Uranus'.
Book 9		'Theory of comets'.
1, 1–9	194	General theory of their perturbations.
2, 10–13	216	Perturbations when comet 'approaches very near to a planet'.
3, 14	229	Their masses, and 'action upon the planets'.
Book 10		'On several subjects relative to the system of the World'.
1, 1–10	231	'On astronomical refractions'; differential equation.
2, 11	277	'On terrestrial refractions'.
3, 12–13	282	Extinction of a planet's light in Earth's atmosphere.
4, 14	289	'On the measure of heights by a barometer'.
5, 15–16	294	'On the descent of bodies which fall from a great height'.
6, 17	306	Special cases of many-body problem.
7, 18–22	313	Motions of planets and comets affected by traversed medium, 'or by the successive transmission of gravity'.
8, 23–24	327	'Supplement to the theories of planets'; Jupiter, Saturn, Moon. [End 347.]
Supp.1, 1–16	65 pp.	'Capillary attraction': shape of meniscus; fluid between tubes or planes; comparison with observations.
Supp.2	80 pp.	'Fundamental equation' again; adhesion of fluids; shape of blob of mercury.

the condition that the angular momentum of the system was maximal along its normal (1#21–22). In 1830 Louis Poinsot was to criticise the analysis for neglecting the areas created by the rotation of satellites about their planets, and of the planets about their own axes.

Book 2 was mainly devoted to the motions of the planets about the Sun; Laplace started with two-body problems, yielding elliptical orbits for planets though more complicated for comets (2#16–39). Then as the 'second approximation' he considered the inter-planetary perturbations and analysed their 'secular inequalities' (that is, the perturbations which did not depend upon the mutual configuration of the relevant heavenly bodies). Although the configuration required only trigonometry in the invariable plane for expression, Newton's inverse square law caused some horrible expressions in the astronomical variables; so in the late 1740s Euler had made the wonderful simplification of converting them into infinite trigonometric series in multiples of the relevant angles [Wilson, 1980]. (They resemble Fourier series but have a quite different theory.) This procedure became normal especially for French astronomy, with Lagrange and then Laplace, who gave the basic details in 2#46–52. One ground for Laplace's support of them seem to have been his belief that periodic forces produce periodic effects (explicitly in 13#1), and therefore needed periodic func-

tions in the mathematics. To solve the system of associated differential equations he often deployed a method of 'successive approximations' (2#40–45).

A further major question was to prove that the planetary system was stable, which then meant that the eccentricities and inclinations of the orbit of each planet were strictly bounded, to that it would neither go out of finite bounds within the ecliptic nor fly out of that plane. By a brilliant transformation of variables Lagrange had tackled this problem in 1778 as the motion of many point-masses (§16); modifying the analysis somewhat to fit the planetary system, Laplace summarised the findings in 2#55–62. Mathematically the task is demonstrate the reality of the latent roots and latent vectors of matrices linked to the eccentricities and inclinations of the planets' orbits; however, not only is this manner of expression historically anachronistic, but the stability problem itself was crucial in the development of the spectral theory in the first place [Hawkins, 1975]. By analysing the quadratic forms that we would associate with these matrices Lagrange and Laplace found profound but not conclusive results.

Another important perturbation was the apparent resonance in the mean motions of Jupiter and Saturn, and of three of Jupiter's satellites. Pioneering the analysis of perturbation terms in powers of eccentricity and/or inclination, Laplace's studies of the mid 1780s had been among his early successes [Wilson, 1985], and he summarised his and Lagrange's findings in 1#65–72.

5 *MÉCANIQUE CÉLESTE*, THE PLANETARY VOLUME 2

In Book 3 Laplace turned to questions concerning the shape of the planets, especially the Earth. He took the potential $\int_B \rho\, dv/r$ of body B with density ρ at point w distant r from the attracted point, showed that it satisfied Laplace's equation, and when set in spherical polar co-ordinates he solved it by spherical harmonics. The language here is modern, and Laplace's presentation (in other terms) is familiar (3#1–17) apart from the names (not even Laplace would have referred to 'Laplace's equation'!); but in fact it was an *early* account of the theory, which he had done much to develop since the 1770s. Again there had been competition, this time mainly from Legendre [Todhunter, 1873, chs. 19–28]; indeed, the 'Legendre functions' in the harmonics were known (following William Whewell) during the 19th century as 'Laplace coefficients'. Laplace used the power-series expansion (assumed to be convergent) and the generating function, orthogonality, and the expansion of "any" function in an infinite series of the functions; the associated function occasionally appeared. The indexes were usually integral, when the functions were polynomial. For some reason (his attitude to Legendre?) he did not use elliptic integrals.

For mechanics Laplace naturally focused upon the nearly spherical spheroid, and handled its ellipticity by means similar to his method of successive approximations (3#33). His first application was to determine the 'figure' of a homogeneous fluid of constant thickness covering it and rotating in equilibrium (3#22–37). Much of the analysis dealt with 'level surfaces' (Colin MacLaurin's name for equipotentials).

To compare his findings with available data for the Earth, Laplace gave statistics a rare airing in the *Mécanique céleste*, refining earlier studies by R.J. Boscovich (3#39–40). According to theory, the length $l(\theta)$ of the meridian for a degree of arc at latitude θ° was

proportional to $\sin 2\theta$; so he formed the error expression

$$E(\theta) := l(\theta) - A\sin^2\theta - l(0), \tag{1}$$

where A was related to ellipticity in a known way. Should the observations be exact, then $E(\theta) = 0$; but life was never so kind. Hence he proposed to take the available data at latitudes θ_j determined from data of range $(n_j/2)^\circ$ each side, and calculate A from (1) by the criteria that

$$\sum_j n_j l(\theta_j) = 0 \quad \text{and} \quad \sum_j \left| n_j l(\theta_j) \right| \quad \text{be minimal.} \tag{2}$$

Comparison with the evidence was not too encouraging (3#41–43).

Book 4 was devoted to the closely related topic of sea-flow, where in earlier work Laplace had pioneered a dynamic analysis, with trigonometric series well to the fore. It rested on distinguishing three different periodicities: one monthly and partly annual, and due to the orbit of the Earth; one diurnal, and caused by its rotation; and one semi-diurnal, largely blamed upon the Moon (4#5–9). Comparison with data again led to discrepancy, especially for the port of Brest, which had been well studied for the length and heights of its tides (4#23); but he discussed in detail the difference between tides in syzygy and in quadrature with Sun and Moon (4#22–42).

In Book 5 of this volume Laplace briefly analysed lunar librations and the motions of its nodes (5#15–19). The main attack on the Moon would come later; the motivation here was to consider effects of the rotation of a body about its centre of gravity, the subject of this Book. He followed with the potential of the 'ring' of Saturn [Cooke, 1984, 66–74]. Relying upon observations that claimed it actually to be two concentric rings, he concluded that each one was a ellipse with small thickness rotating at its own angular velocity, and that its material was not distributed uniformly so that its centre of gravity did not coincide with that of the planet (3#44–46). The latter finding led him later in the volume to analyse the motion of the rings about their centres (5#20–22). Finally here he analysed the (assumedly solid) Earth, where he handled the precession and nutation of the polar axis by means of Euler's equations for the rotation of such a body about a point; he deduced that the effects of the sea as a stratum, and of winds, could be ignored (5#12–14).

6 *MÉCANIQUE CÉLESTE*, THE NUMERICAL BOOKS 6–9

The last two volumes appeared in 1802 and 1805, constituting the 'Second Part' of the work. Laplace dealt with the 'Particular theories' of the motions after the generalities and principles just expounded; but the final Book 10 dealt with various other topics, and so will be noted in the next section.

Table 2 shows the remarkable panoply of inequalities and perturbations that Laplace presented in these Books; they were also the venue for the longest trigonometric series, with coefficients calculated to many decimal places (for example, 6#33 on Jupiter, including values for 1950 as well as for 1750). Often these values were small, and a frequent concern in the analyses was to determine upper bounds for terms containing various powers of eccentricity and/or inclination. Mathematical novelties occurred much less often;

but, for example, he explained the role of 'generating functions' to approximate to the elements of the path of a comet (9#5).

The largest single theoretical effort was given over to lunar theory, to which Book 7 was devoted. Various methods had been introduced during the 18th century to analysis the many perceived perturbations of this nearby object; Laplace principally favoured one due to d'Alembert in which time was set as a function of the Moon's true longitude in the ecliptic, not vice versa. The equations took the form of integral–differential, then unusual (7#1); after finding solutions Laplace desimplified them by introducing knowingly neglected factors such as the effect of the action of the Sun and of the Earth's eccentricity upon the Moon's secular acceleration (7#10, 16). In his analysis of lunar parallax he allowed for the oblateness of the Earth (7#20–21). He claimed good correspondence with certain observational data, such as the lunar perigee (7#16).

Another subject of especial difficulty was the theory of comets. In Book 2 Laplace had conducted a preliminary analysis in which all conic sections were permitted as paths (2#23); now in Book 9 he again approximated by taking the path to be nearly elliptic and using generating functions to effect quadratures. The analysis is exceptionally laborious, a point to be considered in the last section below.

7 *MÉCANIQUE CÉLESTE,* THE MISCELLANEOUS BOOK 10

Some of the material here could have been presented earlier, and may have constituted afterthoughts or late news. For example, Laplace analysed the path of a projectile falling to Earth from a great height; a striking feature of this use of Newton's second law is his allowance for the rotation of the Earth, where he included components of the force named now after his successor G.G. Coriolis (10#15–16). One motivation for this excursion into the stratosphere may have been recent French experience of meteorites; Laplace had wondered if they were rocks detached from the Moon, and around 1800 Biot and S.D. Poisson (1781–1840) had examined the consequences [Grattan-Guinness, 1990, 388–400].

Another speculation concerned one of Newton's greatest mysteries: *how* does that gravitational force pass between bodies? Laplace presumed that 'the successive transmission of gravity' was carried by an elastic aether, and thereby analysable by the usual equations; by making assumptions about the (minute) loss of mass by the Sun caused by the attractions, he found the velocity to be 'about seven millions of times greater than that of light' (10#22).

The major feature of this Book was its attention to physics. This analysis of gravitational involved light from the Sun, and the first and largest part of the Book was an analysis of atmospheric refraction, an exercise in small effects partly motivated by the ever-improving accuracy of astronomical instruments. Adopting a form of Newton's optics, that light was composed of tiny fast-moving bullets, Laplace construed refraction to be caused by interaction with the molecular constituents of the atmosphere by central forces, whose action function f was known only to decline very rapidly as distance from source increased. Such cumulative action C among molecules was expressed as an integral involving f; the motion was analysed as a differential equation with C among the coefficients. The constitution of the refracting atmosphere had also to be considered; Laplace suggested four

density functions of altitude (10#1–7). He contented himself with some special cases and results.

However, much greater ambition attended Laplace's analysis of the use of the barometer. In the Book he only related atmospheric pressure to density (10#14); but soon afterwards he subjected the analysis of the meniscus to an intense molecularist analysis. It appeared in various papers and especially two supplements to this volume, published in 1806 and 1807 and at 145 pages longer than several of the Books. Uniquely in the entire work, the first supplement caused several diagrams.

In the first supplement Laplace explained the meniscus in the capillary tubes of barometers and thermometers in terms of action between the molecules of the fluid contained therein and those in the surrounding glass; in particular, the closer the fluid to the glass, the stronger the attracting force. Using integrals again to express such cumulative action, he found a foul differential equation for the shape of the meniscus, to be solved by successive approximations (supp.1#1–5). He also analysed the shape of fluid trapped between two glass tubes or between planes (supp.1#6–8). Experimental data on these and other cases, some obtained at his request, were again not very encouraging (supp.1#15–16). In the second supplement he re-derived the basic differential equation, considered the capacity of the surface of and fluid to bear weight, and especially studied the shape of a blob of mercury in equilibrium on a horizontal plane.

8 IMMEDIATE INFLUENCE

In these supplements Laplace had moved far away from orthodox celestial mechanics, though still within the general concerns of planetary mechanics; but he had permanently changed the balance of his interests. During the 1800s he and the chemist Claude Berthollet led an unofficial but influential school of physicists and chemists, especially bright young graduates of the *Ecole Polytechnique*, where molecular modelling was the dominant approach [Fox, 1974]. Physics and chemistry dominated over mathematics, but Laplace and his loyal follower Poisson mathematicised phenomena in terms of integrals with integrands containing f [Grattan-Guinness, 1990, ch. 8]. Optics was the most successful area, with Etienne Malus the most notable contributor; heat diffusion also fell under the sway (§26.8).

In addition during the later 1800s, Laplace revived an earlier concern with probability and mathematical statistics. Including a somewhat unpleasant priority dispute over least squares regression mainly between Legendre and C.F. Gauss (1777–1855) around 1810, it culminated with his great treatise of 1812 and the *Exposition*-like summary of 1814 (§24).

However, Laplace still attended to celestial mechanics, especially because of Poisson. Most of Laplace's analysis of perturbations in the *Mécanique céleste* had been executed to the first order in the planetary masses; but he had examined second-order terms occasionally (especially 3#12–18), and in 1807 Poisson examined them in general. The main consequence of the work was 'Lagrange–Poisson brackets' theory of canonical solutions to the equations of motion, which was quickly produced by these two men [Grattan-Guinness, 1990, 371–386]. In a short supplement to volume 3 of the *Mécanique céleste* published in 1808, Laplace praised Poisson's finding that the stability of the planetary system was indeed not endangered by these effects.

Laplace's achievements were prominent in a report on the progress of science since 1789 written by Delambre in 1808 for Emperor Napoléon [Delambre, 1810]; and two years later *Mécanique céleste* won, probably easily, a prize in 1810 of the *Institut de France* for the best work of the decade in astronomy and mechanics. It had also begun to have impact in other countries, especially Britain: not only were translations made or started (as listed above), but British mathematicians had begun to transfer allegiance from Newton's fluxional calculus to Leibniz's differential version (§4) and Lagrange's algebraic alternative (§19), and interest in *Mécanique céleste* was the main single cause [Guicciardini, 1989, pt. 3]. Various commentaries began to appear; Thomas Young gave some *Elementary illustrations* (1821) of Book 1 in the form of a free translation rather swamped by his own opinions; Mary Somerville produced her much appreciated summary *Mechanism of the heavens* (1831); and above all for everybody, including the French, was Bowditch's American translation/edition of volumes 1–4 (1829–1839), still the most informative version.

9 *MÉCANIQUE CÉLESTE*, THE MISCELLANEOUS VOLUME 5

Laplace's molecular physics was *not* followed by more successful practitioners of mathematical physics: Fourier on heat diffusion (§26) from the late 1800s, and A.J. Fresnel in optics from the mid 1810s. For the fourth edition (1814) of the *Exposition* Laplace had added a long new chapter on molecularism; but he omitted it (and two others) from the fifth (1824), promising a separate book which however was never written and maybe not really envisaged. Indeed, at his death he was planning the sixth edition with the omitted chapters reinstated; this decision was fulfilled upon its posthumous appearance in 1835 under the guidance of his son. Confusingly, the *Oeuvres* edition of 1846 has the fifth edition, but the *Oeuvres complètes* of 1884 takes the sixth.

Laplace's interest in celestial mechanics revived in the 1820s, especially when he published the fifth and final volume of the *Mécanique céleste* in six *fascicules* between 1823 and 1825, with a posthumous supplement in 1827. Once again most of the material had appeared over the years in papers. As Table 3 shows, he elaborated upon a wide variety of the concerns of the earlier volumes (though with some inconsistency in the numbering of Sections). The basic approach was left intact, though with additions such as the loss of gravitational attraction between two bodies when a third one interposed (16#6). Some revisions were inspired by others' work; for example, the rotation of the Earth in view of contributions made by Poisson (14#2–3). The interest in heat diffusion showed in a discussion of the cooling and the age of the Earth (11#9–10); and in the velocity of sound, which for Laplace depended upon the specific heats of air (12#7). His increased involvement with probability and mathematical statistics is evident in several forays: for example, the tides, where he used a method similar to that described in section 5 above (13#5), and the probability of the existence of a 'lunar atmospheric tide' (13#ch.6; and supp#5, his last piece of work). He also fulfilled an aim manifest already in the *Exposition* of writing at some length on the history of astronomy; indeed, in 1821 he had pre-published the historical Book 5 of the fifth (1824) edition of the *Exposition* as a separate *Précis.*

The fifth volume of the *Mécanique céleste* made less impact than its predecessors; perhaps one reason was Bowditch's failure to fulfil his intention of translating it. But another cause may lie in the reception of the tendency in celestial mechanics which Laplace

Table 3. Contents of *Mécanique céleste*, volume 5 (1823–1827). The second column gives either the month of the date of publication of a Book or the original first page number of a chapter.

Ch., arts.	**Date/pp.**	**Topics**
Book 11	3.1823	'On the figure and rotation of the Earth'.
1, 1	2	Historical notice of work by 'geometers' on celestial mechanics.
2, 2–7	22	'On the figure of the Earth'.
3, 8	57	'On the axis of rotation of the Earth'.
4, 9–10	72	'On the heat of the Earth' and reduction of the day by cooling.
Book 12	4.1823	'On the attraction and the repulsion of spheres, and on the laws of equilibrium and on the motion of elastic fluids'.
1, 1	87	'Historical notice of researches by geometers on this subject'.
2, 2–6	100	Attraction of spheres, and repulsion of elastic fluids.
3, 7–12	119	Speed of sound, motion of elastic fluids and of aqueous vapour.
Book 13	2.1824	'On the oscillations of fluids which cover the planets'.
1, 1	145	Historical notice, especially the 'ebb and flow of the sea'.
2, 2–4	168	New researches on the theory of seas.
3, 5–11	183	Comparison of analysis with observations of height of seas.
4, 12	213	Ditto with 'hours and intervals of the seas'.
5, 13	221	On partial flows of nearly daily period.
6, –	230	On partial flows depending upon (lunar distance)$^{-4}$.
7, 1–2	237	Flow of the atmosphere; remarks on vol. 1, p. 115.
Book 14	7.1824	'On the motions of the celestial bodies about their centres of gravity'.
1, 1–4	245	Precession of equinoxes: historical notice; formulae for terrestrial equator.
2, 5–6	278	Libration of the Moon: historical notice; remarks.
3, –	288	Rings of Saturn; historical notice.
Book 15	12.1824	'On the motion of planets and of comets'.
1, 1	293	Historical notice.
2, 2–5	328	Varia from Book 2: series; Jupiter-Saturn inequality; orbits of comets.
Book 16	8.1825?	'On the motion of satellites'.
1, 1	349	Lunar motion; historical notice.
2, 2–3	367	'On the lunar theory of Newton'.
3, 4–5	381	On long-period inequalities depending upon the difference between two terrestrial hemispheres; lunar inequalities.
4, 6	401	'On the law of universal attraction'.
5, 7	408	'On the motions of the satellites of Jupiter'; historical notice.
6, 8	415	'Influence of great inequalities of Jupiter on the motions of its satellites.
7, 9	417	'On the satellites of Saturn and of Uranus'. [End 420.]
Supp., 1–8	35 pp. 1827	To Books 11, 13 and 15: series expansions; lunar atmosphere.

favoured. We saw in section 4 that, following Euler and Lagrange, he preferred the use of trigonometric series to express perturbations and other phenomena. After him the practice was continued by several compatriots, especially Biot and Poisson, then G. de Pontécoulant (the author of his own large treatise, *Théorie analytique du système du monde* in four volumes, 1829–1846), and then Urban Leverrier and Charles Delauney. But when one considers, for example, the latter's lunar theory of the 1860s, where the astronomical variables are expressed in series of literally hundreds of terms, and literally hundreds of pages are needed to analyse their respective orders of magnitude—then one may wonder if such 'exactness' (a favourite Laplace word) was the best way forward.

This question seems to have been in German minds. At all events, they adopted an alternative strategy: the phenomena are too complicated to handle 'exactly'; so accept approximation and proffer compact feasible methods. Two years before *Mécanique céleste* appeared Wilhelm Olbers had given a nice means of approximating to the paths of comets which contrasts strongly with Laplace's lucubrations mentioned in section 6. Soon afterwards Gauss used a method of this kind to analyse the motion of the recently discovered minor planets, and then of all heavenly bodies [Gauss, 1809], including his treatment of planetary motions (§23) noted in section 8 in connection with including least squares. In 1834, as a sign of the times for students of celestial mechanics, Bowditch rehearsed both Olbers's and Gauss's contributions in a long appendix to his translation of volume 3.

To sum up, the three reprints of *Mécanique céleste* in the 19th century reflect its high status in France, and to some extent elsewhere; a *locus classicus* for celestial mechanics on a scale unmatched since Newton, and also a valuable source for a cluster of important mathematical theories and methods. Nevertheless, nobody ever translated Laplace's last three volumes into German.

BIBLIOGRAPHY

Biot, J.B. 1802. *Analyse du traité de mécanique céleste par P.S. Laplace*, Paris: Duprat.

Cooke, R. 1984. *The mathematics of Sonya Kovalevskaya*, New York: Springer.

de Prony, R. 1801. *Analyse de l'exposition du système du monde de P.S. Laplace*, Paris: Chollet.

Delambre, J.B.J. 1810. *Rapport historique sur les progrès des sciences mathématiques depuis 1789 . . .*, Paris: *Imprimerie Impériale* [octavo and quarto eds. Octavo repr. Paris: Belin, 1989].

Eisenstaedt, J. 1991. 'De l'influence de la gravitation sur la propagation de la lumière en théorie newtonienne. L'archéologie des trous noirs', *Archive for history of exact sciences, 42*, 315–386.

Fox, R. 1974. 'The rise and fall of Laplacian physics', *Historical studies in the physical sciences, 4*, 81–136.

Gauss, C.F. 1809. *Theoria motus corporum coelestium*, Hamburg: Perthus und Bessor. [Repr. in *Werke*, vol. 7, 1–282.]

Gautier, A. 1817. *Essai historique sur le problème des trois corps*, Paris: Courcier.

Gillispie, C.C. 1998. *Pierre Simon Laplace* (*1749–1827*), Princeton: Princeton University Press.

Grant, R. 1852. *History of physical astronomy, from the earliest ages to the middle of the nineteenth century*, London: Boh. [Repr. New York: Johnson, 1966.]

Grattan-Guinness, I. 1990. *Convolutions in French mathematics, 1800–1840*, 3 vols., Basel: Birkhäuser; Berlin: Deutscher Verlag der Wissenschaften.

Guicciardini, N. 1989. *The development of Newtonian calculus in Britain 1700–1800*, Cambridge: Cambridge University Press.

Hawkins, T.W. 1975. 'Cauchy and the spectral theory of matrices', *Historia mathematica, 2*, 1–29.
Taton, R. and Wilson, C. (eds.) 1995. *The general history of astronomy*, vol. 2, pt. B, Cambridge: Cambridge University Press. [See especially the articles by B. Morando and J. Laskar.]
Todhunter, I. 1861. *A history of the progress of the calculus of variations during the nineteenth century*, Cambridge and London: Macmillan. [Repr. New York: Chelsea, 1961.
Todhunter, I. 1873. *A history of the mathematical theories of attraction and figure of the earth* ..., 2 vols., London: Macmillan. [Repr. New York: Dover, 1962.]
Wilson, C. 1980. 'Perturbation and solar tables from Lacaille to Delambre', *Archive for history of exact sciences, 22*, 53–188, 189–304.
Wilson, C. 1985. 'The great inequality of Jupiter and Saturn: from Kepler to Laplace', *Ibidem, 33*, 15–290.

CHAPTER 19

JOSEPH LOUIS LAGRANGE, *THÉORIE DES FONCTIONS ANALYTIQUES*, FIRST EDITION (1797)

Craig G. Fraser

In this volume, based upon his first teaching at the *Ecole Polytechnique*, Lagrange both popularised and extended his view that the differential and integral calculus could be based solely on assuming the Taylor expansion of a function in an infinite power series and on algebraic manipulations thereafter. He also made some applications to problems in geometry and mechanics.

First publication. Paris: L'Imprimerie de la République, 1797. 277 pages.

Reprint. As *Journal de l'Ecole Polytechnique, cahier 9, 3* (1801), 1–277.

Later editions. 2nd 1813, 3rd 1847. Both Paris: Bachelier. Also as Lagrange *Oeuvres*, vol. 9, Paris: Gauthiers–Villars, 1881.

Portuguese translation. *Theorica das funções analyticas* (trans. M.J. Noguiera da Gama), 2 vols., Lisbon: J.C.P. da Silva, 1798.

German translations. 1) Of 1st ed.: *Theorie der analytischen Functionen* (trans. J.P. Gruson), 2 vols., Berlin: Lagarde, 1798–1799. 2) Of 2nd ed.: *Theorie der analytischen Functionen* (trans. A.L. Crelle), Berlin: Reimer, 1823.

Related articles: Leibniz (§4), MacLaurin (§10), Euler (§12–§14), Lagrange on mechanics (§16), Lacroix (§20), Cauchy on real-variable analysis (§25).

1 INTRODUCTION

At the end of the 18th century Joseph Louis Lagrange (1736–1813) published a book in which he developed a systematic foundation of the calculus, his *Théorie des fonctions analytiques* (1797). Parts of it were further developed in his *Leçons sur le calcul des functions* (1801; revised edition 1806).

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

By 1790 a critical attitude had developed both within mathematics and within general scientific culture. As early as 1734 Bishop George Berkeley in his work *The Analyst* had called attention to what he perceived as logical weaknesses in the reasonings of the calculus arising from the employment of infinitely small quantities (§8). Although his critique was somewhat lacking in mathematical cogency, it at least stimulated writers in Britain and the Continent to explain more carefully the basic rules of the calculus. In the 1780s a growing interest in the foundations of analysis was reflected in the decisions of the academies of Berlin and Saint Petersburg to devote prize competitions to the metaphysics of the calculus and the nature of the infinite. In philosophy Immanuel Kant's *Kritik der reinen Vernunft* (1787) set forth a penetrating study of mathematical knowledge and initiated a new critical movement in the philosophy of science.

The contents of the *Théorie des fonctions analytiques* is summarised in Table 1. The book was divided in three parts, the first part devoted to analysis and the second and third

Table 1. Contents of Lagrange's book. The original pagination is given.

Page	**Topics**
Part One	*Exposition of the theory, with its principal uses in analysis.*
1	Preliminaries, series developments, and derived functions.
15	Series expansions for algebraic and transcendental functions.
28	Composite functions, exceptional cases.
41	Expression for the remainder.
50	Equations among derived functions.
80	Primitive functions.
91	Functions of several variables.
Part Two	*Application of the theory to geometry and to mechanics.*
	Application to geometry.
117	Theory of contacts.
147	Developable curves.
150	Maxima and minima of a function of a single variable.
155	Areas and path-lengths.
161	Differential geometry of surfaces.
187	Maxima and minima.
200	Method of variations.
	Application to mechanics.
223	Speed and acceleration.
232	Particle dynamics.
241	Motion in a resisting medium.
251	Constrained motion.
256	Conservation theorems.
271	Impact of bodies, machine performance. [End 277.]

parts devoted to geometry and mechanics. The *Leçons sur le calcul des fonctions* concentrated almost exclusively on analysis, and included a detailed account of the calculus of variations. The material in both books originated in lectures that Lagrange delivered at the *Ecole Polytechnique*. He wrote them when he was in his sixties, still an active mathematician but certainly past his prime creative period of scientific research.

Although Lagrange's books appeared at the dawn of the new century, they encapsulated the prevailing understanding of analysis, refining conceptions that had been set forth by Leonhard Euler (1707–1783) in his textbooks of the 1740s and 1750s (§13, §14). Lagrange's fundamental axiom involving the Taylor-series expansion of a function originated in a memoir he published in 1774 in the proceedings of the Berlin Academy. Near the beginning of the *Théorie*, he stated that Louis Arbogast had submitted a detailed memoir to the Paris Academy developing these ideas. The memoir was never published, although Arbogast discussed it in his book [1800]; because it did not appear, and because Lagrange himself happened to be involved in a study of the general principles of analysis as a result of 'particular circumstances' (presumably his teaching duties), he decided to write a treatise generalizing and extending his earlier ideas.

Part One of the *Théorie* begins with some historical matters and examines the basic expansion of a function as a Taylor power series. There is considerable discussion of values where the expansion may fail, and a derivation of such well-known results as l'Hôpital's rule. Lagrange then turned to methods of approximation and an estimation of the remainder in the Taylor series, followed by a study of differential equations, singular solutions and series methods, as well as multi-variable calculus and partial differential equations. He outlined and supplemented topics explored in some detail in memoirs of the 1760s and 1770s.

Part Two on geometry opens with an investigation of the geometry of curves. Here Lagrange examined in detail the properties that must hold at a point where two curves come into contact—the relationships between their tangents and osculating circles. Corresponding questions concerning surfaces are also investigated, and Lagrange referred to Gaspard Monge's memoirs on this subject in the *Académie des Sciences*. He derived some standard results on the quadrature and rectification of curves. The theory of maxima and minima in the ordinary calculus, a topic Lagrange suggested could be understood independently of geometry as part of analysis, is taken up. Also covered are basic results in the calculus of variations, including an important theorem of Adrien-Marie Legendre in the theory of sufficiency. The topic of the calculus of variations was treated on an analytical level much more extensively in the *Leçons*.

The third part on dynamics is somewhat anticlimactic, given the publication nine years earlier of his major work *Méchanique analitique* (§16). In this part Lagrange presented a rather kinematically-oriented investigation of particle dynamics, including a detailed discussion of the Newtonian problem of motion in a resisting medium. He also derived the standard conservation laws of momentum, angular momentum and live forces. The book closes with an examination of the equation of live forces as it applies to problems of elastic impact and machine performance.

In our account of the *Théorie* we will concentrate on some of the major original contributions of this work: the formulation of a coherent foundation for analysis; Lagrange's conception of theorem-proving in analysis; his derivation of what is today called the Lagrange remainder in the Taylor expansion of a function; his formulation of the multiplier

rule in the calculus and calculus of variations; and his account of sufficiency questions in the calculus of variations. Pages of this book, and also from the *Leçons*, are cited from the *Oeuvres* edition.

2 ALGEBRAIC ANALYSIS AND THE FUNCTION CONCEPT

The full title of the *Théorie* explains its purpose: 'Theory of analytical functions containing the principles of the differential calculus disengaged from all consideration of infinitesimals, vanishing limits or fluxions and reduced to the algebraic analysis of finite quantities'. Lagrange's goal was to develop an algebraic basis for the calculus that made no reference to infinitely small magnitudes or intuitive geometrical and mechanical notions. In a treatise on numerical equations published in 1798 he set forth clearly his conception of algebra [1798, 14–15]:

> [Algebra's] object is not to find particular values of the quantities that are sought, but the system of operations to be performed on the given quantities in order to derive from them the values of the quantities that are sought. The tableau of these operations represented by algebraic characters is what in algebra is called a *formula*, and when one quantity depends on other quantities, in such a way that it can be expressed by a formula which contains these quantities, we say then that it is a *function* of these same quantities.

Lagrange used the term 'algebraic analysis' to designate the part of mathematics that results when algebra is enlarged to include calculus-related methods and functions. The central object here was the concept of an analytical function. Such a function $y = f(x)$ is given by a single analytical expression, constructed from variables and constants using the operations of analysis. The relation between y and x is indicated by the series of operations schematized in $f(x)$. The latter possesses a well-defined, unchanging algebraic form that distinguishes it from other functions and determines its properties.

The idea behind Lagrange's theory was to take any function $f(x)$ and expand it in a power series about x:

$$f(x+i) = f(x) + pi + qi^2 + ri^3 + si^4 + \cdots. \tag{1}$$

The 'derived function' or derivative $f'(x)$ of $f(x)$ is defined to be the coefficient $p(x)$ of the linear term in this expansion. $f'(x)$ is a new function of x with a well-defined algebraic form, different from but related to the form of the original function $f(x)$. Note that this conception is very different from that of the modern calculus, in which the derivative of $f(x)$ is defined at each value of x by a limit process. In the modern calculus the relationship of the derivative to its parent function is specified in terms of correspondences defined in a definite way at each value of the numerical continuum.

Lagrange's understanding of derived functions was revealed in his discussion in the *Leçons* of the method of finite increments. This method was of historical interest in the background to his programme. Brook Taylor's original derivation in 1715 of Taylor's theorem was based on a passage to the limit of an interpolation formula involving finite increments. Lagrange wished to distinguish clearly between an approach to the foundation of

the calculus that uses finite increments and his own quite different theory of derived functions. In taking finite increments, he noted, one considers the difference $f(x_{n+1}) - f(x_n)$ of the same function $f(x)$ at two successive values of the independent argument. In the differential calculus the object Lagrange referred to as the derived function was traditionally obtained by letting $dx = x_{n+1} - x_n$ be infinitesimal, setting $dy = f(x_{n+1}) - f(x_n)$, dividing dy by dx, and neglecting infinitesimal quantities in the resulting reduced expression for dy/dx. Although this process leads to the same result as Lagrange's theory, the connection it presumes between the method of finite increments and the calculus obscures a more fundamental difference between these subjects: in taking $\Delta y = f(x_{n+1}) - f(x_n)$ we are dealing with one and the same function $f(x)$; in taking the derived function we are passing to a new function $f'(x)$ with a new algebraic form. Lagrange explained this point as follows [1806, 270, 279]:

> [...] the passage from the finite to the infinite requires always a sort of leap, more or less forced, which breaks the law of continuity and changes the form of functions.
> [...] in the supposed passage from the finite to the infinitely small, functions actually change in nature, and [...] dy/dx, which is used in the differential Calculus, is essentially a different function from the function y, whereas as long as the difference dx has any given value, as small as we may wish, this quantity is only the difference of two functions of the same form; from this we see that, if the passage from the finite to the infinitely small may be admitted as a mechanical means of calculation, it is unable to make known the nature of differential equations, which consists in the relations they give between primitive functions and their derivatives.

In Lagrange's conception of analysis, one is given a universe of functions, each expressed by a formula $y = f(x)$ and consisting of a single analytical expression involving variables, constants and algebraic and transcendental operations. During the 18th century such functions were called continuous, and the *Théorie* is devoted exclusively to functions that are continuous in this sense. (Mathematicians were aware of the possibility of other sorts of functions, but alternate definitions never caught on.) Such functions were naturally suited to the usual application of calculus to geometrical curves. In studying the curve the calculus is concerned with the connection between local behaviour, or slope, and global behaviour, or area and path-length. If the curve is represented by a function $y = f(x)$ given by a single analytical expression then the relation between x and y is permanently established in the form of f. Local and global behaviour become identified in this functional relation.

It is also necessary to call attention to the place of infinite series in Lagrange's system of analysis. Each function has the property that it may be expanded as the power series (1). Nevertheless, an infinite series as such is never defined to be a function. The logical concept of an infinite series as a functional object defined *a priori* with respect to some criterion such as convergence or summability was foreign to 18th-century analysis. Series expansions were understood as a tool for obtaining the derivative, or a way of representing functions that were already given.

For the 18th-century analyst, functions are things that are given 'out there', in the same way that the natural scientist studies plants, insects or minerals, given in nature. As a gen-

eral rule, such functions are very well-behaved, except possibly at a few isolated exceptional values. It is unhelpful to view Lagrange's theory in terms of modern concepts (arithmetical continuity, differentiability, continuity of derivatives and so on), because he did not understand the subject in this way.

3 THEOREMS OF ANALYSIS

3.1 Expansions

Lagrange was aware that the expansion of a function as the series (1) may fail at particular values of x, and he discussed this point at some length in the *Théorie*. He reasoned that the expansion of $f(x+i)$ can contain no fractional powers of i. He illustrated this conclusion by means of the example $f(x) = \sqrt{x}$. Suppose indeed that we had a relation of the following form for the expansion of $\sqrt{(x+i)}$:

$$\sqrt{x+i} = \sqrt{x} + ip + i^2q + i^3r + \cdots + i^{m/n}. \tag{2}$$

This equation establishes a relation of equality between the 2-valued function on the left side, and the $2n$-valued function on the right side, a result that is evidently absurd. Hence it must be the case that the powers of i in the expansion of $\sqrt{(x+i)}$ are all integral.

Lagrange noted that the 'generality' and 'rigour' of this argument require that x be indeterminate (p. 8: it is interesting that he associates generality and rigour, a point of view characteristic of 18th-century algebraic analysis). In particular cases such as $x = 0$ we will have fractional powers of i, but this arises because certain formal features of the function—in the case at hand the radical $\sqrt{x}$—disappear at $x = 0$.

3.2 Taylor's theorem

Lagrange's understanding of what it meant to prove a theorem of analysis differed from the understanding which developed in later analysis and which is customary today. To prove a theorem was to establish its validity on the basis of the general formal properties of the relations, functions, and formulae in question. The essence of the result was contained in its general correctness, rather than in any considerations about what might happen at particular numerical values of the variables.

The derived function $f'(x)$ is the coefficient of the linear term in the expansion of $f(x+i)$ as a power series in i. By definition, the second derived function $f''(x)$ is the coefficient of i in the expansion of $f'(x+i)$, the third derived function $f'''(x)$ is the coefficient of i in the expansion of $f''(x+i)$, and so on.

In art. 16 Lagrange related the coefficients $q, r, s, \ldots$ in (1) to the higher-order derived functions $f''(x)$, $f'''(x)$, $f^{(\mathrm{iv})}(x), \ldots$. If we replace i by $i+o$ in (1) we obtain

$$\begin{aligned} f(x+i+o) &= f(x) + (i+o)p = (i+o)^2q + (i+o)^3r + (i+o)^4s + \cdots \\ &= f(x) + ip + i^2q + i^3r + i^4s + \cdots \\ &\quad + op + 2ioq + 3i^2or + 4i^3os + \cdots. \end{aligned} \tag{3}$$

Suppose now the we replace x by $x+o$. $f(x)$, p, q, r then become

$$f(x)+op+\cdots, \quad p+op'+\cdots, \quad q+oq'+\cdots, \quad r+or'+\cdots. \tag{4}$$

If we next increase $x+o$ by i we obtain (using $x+i+o=(x+o)+i$)

$$\begin{aligned} f(x+i+o) = f(x)+op+\cdots+i(p+op'+\cdots)+i^2(q+oq'+\cdots) \\ +i^3(r+or'+\cdots)+\cdots. \end{aligned} \tag{5}$$

Equating (3) and (5) we obtain

$$q=\frac{1}{2}p', \quad r=\frac{1}{3}q', \quad s=\frac{1}{4}r', \quad \ldots. \tag{6}$$

The derived functions $f'(x)$, $f''(x)$, $f'''(x)$, ... are the coefficients of i in the expansions of $f(x+i)$, $f'(x+i)$, $f''(x+i)$, ... Hence

$$q=\frac{1}{2}f''(x), \quad r=\frac{1}{2\cdot 3}f'''(x), \quad s=\frac{1}{2\cdot 3\cdot 4}f^{(\mathrm{iv})}(x), \quad \ldots. \tag{7}$$

Thus the series (1) becomes

$$f(x+i)=f(x)+if'(x)+\frac{i^2}{2}f''(x)+\frac{i^3}{2\cdot 3}f'''(x)+\frac{i^4}{2\cdot 3\cdot 4}f^{(\mathrm{iv})}+\cdots, \tag{8}$$

which is the Taylor series for $f(x+i)$.

3.3 The theorem on the equality of mixed partial derived functions

In art. 86 Lagrange considered a function $f(x,y)$ of the two variables x and y. He observed that $f(x+i, y+o)$ can be expanded in two ways. First, we expand $f(x+i, y+o)$ with respect to i, and then expand the expression which results with respect to o. The expansion for $f(x+i, y+o)$ obtained in this way is presented at the top of p. 93:

$$\begin{aligned} f(x+i,y+o) = f(x,y)+if'(x,y)+of_{\prime}(x,y)+\frac{i^2}{2}f''(x,y)+iof'_{,}(x,y) \\ +\frac{o^2}{2}f_{\prime\prime}(x,y)+\frac{i^3}{2\cdot 3}f'''(x,y)+\frac{i^2o}{2}f''_{,}(x,y) \\ +\frac{io^2}{2}f'_{,\prime}(x,y)+\frac{o^3}{2\cdot 3}f_{\prime\prime\prime}(x,y)+\&c. \end{aligned} \tag{9}$$

One of the terms in this expansion is $f'_{\prime}(x,y)$, where the superscript prime denotes partial differentiation with respect to x and the subscript prime denotes partial differentiation with respect to y, and where the differentiation occurs first with respect to x and second with respect to y.

However, we could also expand $f(x+i, y+o)$ with respect to o, and then expand the expression which results with respect to i. In the expansion obtained in this way, we again

have the term $f'_{,}(x, y)$, except that here the partial differentiation occurs first with respect to y and second with respect to x. By equating the two series expansions for $f(x+i, y+o)$ we are able to deduce that the two quantities $f'_{,}(x, y)$ are equal. In modern notation we have $\partial^2 f/\partial x \partial y = \partial^2 f/\partial y \partial x$. Lagrange's notation system does not allow one to indicate the order of differentiation, but fortunately the order does not matter.

This result evidently applies to all functions $f(x, y)$ for all ranges of the variables x and y, except possibly at isolated exceptional values. Lagrange considered a couple of examples. Suppose $f(x, y) = x\sqrt{(2xy + y^2)}$. If we differentiate f with respect to x and then with respect to y we obtain

$$\frac{x+y}{\sqrt{2xy+y^2}} + \frac{x^2 y}{(2xy+y^2)^{3/2}}. \tag{10}$$

However, if we differentiate f with respect to y and then with respect to x we have

$$\frac{2x+y}{\sqrt{2xy+y^2}} - \frac{(x^2+xy)y}{(2xy+y^2)^{3/2}}. \tag{11}$$

Although these two expressions appear to be different, it is not difficult to see that both reduce to the one and the same expression

$$\frac{3x^2y + 3xy^2 + y^3}{(2xy+y^2)^{3/2}}. \tag{12}$$

Lagrange supplied a second example to provide further confirmation of his theorem.

4 METHODS OF APPROXIMATION

4.1 Lagrange's form of the remainder

In arts. 45–53 Lagrange developed results that belong to the core of any modern course in real analysis. Indeed, it is likely that this part of the treatise influenced Augustin-Louis Cauchy when he wrote his famous textbooks initiating modern analysis 25 years later (§25). However, for Lagrange the results in question were not fundamental: they did not belong to the foundation of the subject. His purpose rather was essentially practical, to obtain a result that would be useful in the approximation of functions. Thus he derived an expression for the remainder in the Taylor series when the series is terminated after the nth term. The result allowed for a general method for approximating functions by obtaining the bounds on the error committed if one approximates a function by the first n terms of its Taylor expansion.

Lagrange first proved the following lemma. If $f'(x)$ is positive and finite throughout the interval $a \leqslant x \leqslant b$, then the primitive function $f(x)$ satisfies the inequality $f(b) - f(a) \geqslant 0$. Consider the expansion

$$f(z+i) = f(z) + if'(z) + \frac{i^2}{2} f''(z) + \&c. \tag{13}$$

For sufficiently small i, the linear term in the expansion on the right side will dominate the sum of the remaining terms. Thus if $f'(z)$ is positive, and i is taken to be a sufficiently small positive quantity, it follows that $f(z+i) - f(z)$ will be positive. Consider the succession of values $a, a+i, a+2i, \ldots, a+ni$. By assumption, $f'(a+i)$, $f'(a+2i), \ldots, f'(a+ni)$ are positive. Thus if i is taken positive and small enough, each of the quantities $f(a+i) - f(a)$, $f(a+2i) - f(a+i), \ldots, f(a+(n+1)i) - f(a+ni)$ will be positive. (Lagrange is evidently assuming a uniformity property with respect to $f(z+i) - f(z)$.) If we let $a + (n+1)i = b$ and add together all of the quantities, it follows that $f(b) - f(a) \geqslant 0$. Hence the lemma is proved.

Lagrange explicitly stated the condition that the derived function $f'(x)$ be finite on the given interval because it was clear from examples that the lemma fails otherwise. In the *Leçons* he cited the example $y = 1/(a-z) - 1/a$ $(a > 0)$ [Ovaert, 1976, 222]. The derived function is $1/(a-z)^2$, which is positive everywhere. Nevertheless, for the interval $[0, b]$ $(b > a)$, it is clear that $f(b) - f(0)$ is negative. However, in this example the derived function is infinite at $x = a$, and so the conditions of the lemma do not hold.

We turn now to Lagrange's derivation of the remainder in the Taylor power series. He first introduced a second variable z and wrote $x = (x - xz) + xz$. Series (8) becomes

$$f(x) = f(x-xz) + xzf'(x-xz) + \frac{x^2z^2}{2\cdot 1}f''(x-xz) + \frac{x^3z^3}{3\cdot 2\cdot 1}f'''(x-xz) + \cdots. \quad (14)$$

We rewrite (14) in the form

$$f(x) = f(x-xz) + xP(x,z). \quad (15)$$

If we differentiate (15) with respect to z we obtain

$$0 = -xf'(x-xz) + xP'(x,z), \quad (16)$$

so that

$$P'(x,z) = f'(x-xz). \quad (17)$$

Suppose that z belongs to the interval $[a, b]$, $a \geqslant 0$. Let N and M be the maximum and minimum values of $P'(x,z)$ on this interval. We have the inequalities

$$N \leqslant P'(x,z) \leqslant M, \quad a \leqslant z \leqslant b. \quad (18)$$

It follows that $P'(x,z) - N \geqslant 0$ and $M - P'(x,z) \geqslant 0$ for $a \leqslant z \leqslant b$. Applying the above lemma to the functions $P - N$ and $M - P$ we obtain

$$P(x,a) + N(b-a) \leqslant P(x,b) \leqslant P(x,a) + M(b-a). \quad (19)$$

Now $P(x,z) = 0$ if $z = 0$. Setting $a = 0$ and $b = 1$ in (19) we obtain

$$N \leqslant P(x,1) \leqslant M. \quad (20)$$

As z goes from 0 to 1, $(x - xz)$ goes from x to 0. From (19) and (20) it follows that $f'(x - xz)$ takes on all values between N and M. (Lagrange is assuming here that $f'(x - xz)$

satisfies an intermediate-value property.) Hence for some u with $0 \leqslant u \leqslant x$ we have, by (20),

$$P(x, 1) = f'(u). \tag{21}$$

Hence the original series (8) may be written for $z = 1$ as

$$f(x) = f(0) + xf'(u), \quad 0 \leqslant u \leqslant x. \tag{22}$$

(22) expresses what is today called 'the mean-value theorem'.

Let us now write (14) in the form

$$f(x) = f(x - xz) + xzf'(x - xz) + x^2 Q(x, z). \tag{23}$$

By differentiating each side of (23) with respect to z we easily deduce that

$$Q'(x, z) = zf''(x - xz). \tag{24}$$

Let N_1 and M_1 be the minimum and maximum values of $f''(x - xz)$ for $a \leqslant z \leqslant b$:

$$N_1 \leqslant f''(x - xz) \leqslant M_1. \tag{25}$$

Since $z \geqslant a \geqslant 0$, we have

$$zN_1 \leqslant zf''(x - xz) \leqslant zM_1, \quad \text{or} \quad zN_1 \leqslant Q'(x, z) \leqslant zM_1. \tag{26}$$

From the lemma we conclude that

$$Q(x, a) + \frac{N_1(b^2 - a^2)}{2} \leqslant Q(x, b) \leqslant Q(x, a) + \frac{M_1(b^2 - a^2)}{2}. \tag{27}$$

Setting $a = 0$ and $b = 1$ in (27) there follows

$$\frac{N_1}{2} \leqslant Q(x, 1) \leqslant \frac{M_1}{2}. \tag{28}$$

It is clear from (25) and (28) that $Q(x, 1) = f''(u)/2$ for some $u \in [0, x]$. Hence for $z = 1$ series (8) becomes

$$f(x) = f(0) + xf'(0) + \frac{x^2}{2 \cdot 1} f''(u), \quad 0 \leqslant u \leqslant x. \tag{29}$$

Lagrange proceeded to extend the reasoning used to obtain (22) and (29) to derive the equation

$$f(x) = f(0) + xf'(0) + \frac{x^2}{2 \cdot 1} f''(0) + \frac{x^3}{3 \cdot 2 \cdot 1} f'''(u), \quad 0 \leqslant u \leqslant x. \tag{30}$$

As he indicated on p. 49, equations (20), (29) and (30) may be iterated to obtain expressions for $f(x)$ involving derivatives of f of any order evaluated at u for $0 \leqslant u \leqslant x$. He observed

that there results 'a theorem which is new and remarkable for its simplicity and generality'. The theorem gives what is today called 'the Lagrange form' of the remainder in the Taylor series.

By taking the function $g(x) = f(x+z)$ and applying the preceding result to $g(x)$ we obtain immediately

$$f(z+x) = f(z) + xf'(u) = f(z) + xf'(z) + \frac{x^2}{2}f''(u)$$
$$= f(z) + xf'(z) + \frac{x^2}{2}f''(z) + \frac{x^3}{2\cdot 3}f'''(u), \tag{31}$$

&c., where $0 \leqslant u \leqslant x$.

Lagrange called attention to the importance of (31) for methods of approximation and emphasized its utility in geometrical and mechanical problems. Although he gave no examples, the usefulness of (31) is evident in one of the functions that he introduced earlier, the exponential function $f(x) = e^x$. We use it to approximate $f(1) = e$, following the account in [Courant, 1937, 326–327]. For $z = 0$ and $x = 1$ (31) becomes

$$e = 1 + 1 + \frac{1}{2!} + \frac{1}{3!} + \cdots + \frac{1}{n!} + \frac{e^u}{(n+1)!}, \tag{32}$$

where $0 \leqslant u \leqslant 1$. We have

$$e = 1 + 1 + 1/2! + 1/3! + 1/4! + \cdots < 1 + 1 + 1/2 + 1/2^2 + 1/2^3 + \cdots$$
$$= 1 + 2 = 3, \quad \text{or} \quad e < 3. \tag{33}$$

Hence the error committed in neglecting the remainder term in (32) will be less than $3/(n+1)!$ To obtain an approximation of e with an error smaller $1/10{,}000$, we observe that $8! > 30{,}000$, and arrive at the estimate

$$e \approx 1 + 1 + \frac{1}{2} + \frac{1}{3\cdot 2} + \frac{1}{4\cdot 3\cdot 2} + \frac{1}{5\cdot 4\cdot 3\cdot 2} + \frac{1}{6\cdot 5\cdot 4\cdot 3\cdot 2} + \frac{1}{7\cdot 6\cdot 5\cdot 4\cdot 3\cdot 2}$$
$$= 2.71822. \tag{34}$$

4.2 Mean-value theorem

Lagrange obtained equation (21), the mean-value theorem (a term he never used), from his fundamental axiom concerning the expansion of a function in a Taylor series. In modern analysis this result is derived in a different way from the basic properties of continuous and differentiable functions. Today the mean-value theorem is a cornerstone of the foundation of real analysis. To prove a theorem is to establish its correctness for each value of the functional variable. The law of the mean is used in theorem-proving in order to take whatever property or relation that is under consideration and localize it a given number. A typical application of this law is found in the modern proof of the theorem on the equality of mixed partial derivatives (discussed above in section 3.3). One takes a point of the

plane, establishes the equality in question for finite increments, and then extends this result to derivatives using the law of the mean. Lagrange's reasoning involved very different conceptions and assumptions, and indicates the large difference between his perspective and the modern one.

There is in the *Théorie* one place where Lagrange's derivation resembles the proof-structure of modern analysis. This occurs in art. 134, in the second part of the book devoted to the application of analysis to geometry. Consider a function $y = f(x)$, and suppose that y and x are the coordinates of a curve in a rectangular coordinate system. Let $F(x)$ be the area under the curve from $x = 0$ to $x = x$. Using the mean-value theorem and some reasoning involving inequalities, Lagrange was able to show that $F(x)$ is a primitive function for $f(x)$. Of course, this is the result known today as the fundamental theorem of calculus and Lagrange's proof is not dissimilar to the reasoning involved in a modern derivation. It is nevertheless important to appreciate the distinctiveness of his approach. First, the result is not fundamental nor is it even a theorem of analysis, but an application of analysis to geometry. Second, the fundamental notion for Lagrange is always the primitive as an antiderivative, rather than the integral as a sum or an area.

5 MULTIPLIER RULE

5.1 Mechanics

The idea for the multiplier rule seems to have originated in an interesting way in Lagrange's study of statics in Part One of the *Méchanique analitique* [1788, 45–49]; compare §16.3. Suppose that we are given a system of bodies or points, with coordinates x, y, z; x', y', z'; and so on. The system is subject to the constraints or conditions $L = 0$, $M = 0$, $N = 0$, and so on. External forces act on each of the points of the system. According to the principle of virtual velocities (what is know today as virtual displacements or virtual work), equilibrium will subsist if the following equation holds:

$$P\,dp + Q\,dq + R\,dr + \&c. = 0. \tag{35}$$

Here $P\,dp$ is the virtual work (which he called 'moment') that results when the force P acts on the point x, y, z with a corresponding virtual displacement dp. Similar interpretations hold for $Q\,dq$, $R\,dr$ and so on. On way to proceed to a solution is the following. We set $dL = 0$, $dM = 0$, $dN = 0$ and use these relations to eliminate some of the differentials dx, dy, dz; dx', dy', dz'; ...(as many differentials are eliminated as there are constraints $dL = 0$). If we substitute for the eliminated differentials in (35), we will obtain a relation in which each of the resulting differentials may be regarded as independent. By equating to zero the coefficients of these differentials, we obtain the desired equations of equilibrium.

It may well be that the required elimination of differentials is not that easy to carry out. The method of multipliers provides an alternative way of deriving the conditions of equilibrium. We multiply $dL = 0$ by the indeterminate quantity λ, dM by the quantity μ, dN by the quantity ν, and so on. Lagrange asserted that 'it is not difficult to prove by the theory of the elimination of linear equations' that the general equation of equilibrium will become

$$P\,dp + Q\,dq + R\,dr + \mu\,dM + \nu\,dN + \&c. = 0. \tag{36}$$

We now equate the coefficient of each of the differentials in (36) to zero; the resulting equations, in combination with the constraints $dL = dM = dN = \cdots = 0$, allow us to eliminate the multipliers and arrive at the desired equilibrium conditions.

Lagrange provided a rather natural physical interpretation of the multiplier constants λ appearing in (36) [1788, 48–49]. The effect of the constraint is to produce a force that acts on the point, producing the increment of virtual work or moment $\lambda\, dL$. The moments due to the constraints $L = 0$ balance the moments resulting from the external forces P. Thus the system can be regarded as subject to the constraints $L = 0$, or alternatively it can be regarded as entirely free with the constraints replaced by the forces to which they gives rise. In the latter case we obtain the free equation (36). According to Lagrange, this interpretation provides the 'metaphysical reason' why the addition of the terms $\lambda\, dL$ to the left side of (35) allows one to treat the system as entirely free in (36)—indeed, 'it is in this that the spirit of the method [of multipliers] consists'. Thus the method is justified in a natural way using physical considerations, in contrast to the analytical approach usually associated with Lagrange's mathematics [Fraser, 1981, 263–267].

5.2 Calculus

In arts. 131–184 of the *Théorie*, Lagrange moved from the domain of mechanics to analysis in his exposition of methods of maxima and minima in the calculus. Although this investigation occurred in the part of the book devoted to the applications of analysis to geometry, he noted the independence of the basic problem of optimization from considerations of curves (pp. 150–151).

In art. 167 Lagrange took up the problem of maximizing or minimizing a function of the form $f(x, y, z, \ldots)$ of any number of variables that are subject to the constraint $\phi(x, y, z, \ldots) = 0$. If we increase x, y, z and so on by the small increments p, q and r we obtain from the constraint $\phi = 0$ the relation

$$p\phi'(x) + q\phi'(y) + r\phi'(z) + \&c. + \textit{higher-order terms} = 0, \tag{37}$$

where $\phi'(x)$, $\phi'(y)$ and $\phi'(z)$ denote as usual the partial derivatives $\partial\phi/\partial x$, $\partial\phi/\partial y$ and $\partial\phi/\partial z$ of ϕ with respect to x, y and z. To arrive at the equations of maxima and minima we can neglect the higher-order terms in (37). Because f is a maximum or minimum we have as well the condition

$$pf'(x) + qf'(y) + rf'(z) + \&c. = 0. \tag{38}$$

One solution would be to solve (37) for p in terms of $q, r, \ldots$, substitute the resulting expression for p in (38), and equate to zero the coefficients of $q, r, \ldots$. Another solution is obtained by multiplying (37) by the multiplier a and adding the resulting expression to (38). In the equation which results all of the variables $x, y, z, \ldots$ may be regarded as free, and the coefficients of each of the $p, q, r, \ldots$ may be set equal to zero:

$$f'(x) + a\phi'(x) = 0, \quad f'(y) + a\phi'(y) = 0, \quad f'(z) + a\phi'(z) = 0, \quad \&c. \tag{39}$$

(39) together with the constraint $\phi = 0$ provide the equations of solution, allowing us to determine the desired maximizing or minimizing values of a, x and y.

Lagrange did not state explicitly the precise reasoning by which he arrived at (39), but it seems to have developed along the following lines. Multiply (37) by a and add to (38) to obtain:

$$p\big(f'(x)+a\phi'(x)\big)+q\big(f'(y)+a\phi'(y)\big)+r\big(f'(z)+a\phi'(z)\big)+\&c.=0 \qquad (40)$$

In (40), define a so that $f'(x)+a\phi(x)=0$. The first term in (40) then disappears, and we can assume that the remaining variables $y, z, \ldots$ may be varied arbitrarily and independently. Hence we obtain the equations

$$f'(y)+a\phi(y)=0, \qquad f'(z)+a\phi(z)=0, \qquad \ldots . \qquad (41)$$

There are, as is well known today, geometrical ways of justifying the multiplier rule. Thus it is immediately evident (to anyone, for example, who has drawn a trail line on a topographical map) that the maximum or minimum of $f(x, y)$ along the path $\phi(x, y)=0$ will occur when this path runs parallel to a member of the family of contours or level curves $f=$ constant. At the point where this is the case the unit normals to the two curves coincide, and so we obtain equations (39). Such geometrical reasoning was not used by Lagrange, whose approach in the *Théorie* was analytical throughout.

5.3 Calculus of variations

In the ordinary calculus it would be possible to do without the multiplier rule, this rule being a useful but finally unessential technique of solution. In the calculus of variations the situation is very different: in problems of constrained optimization, where the side conditions are differential equations, the multiplier rule is the only general method for obtaining the variational equations. The situation in the ordinary calculus is similar to that of the calculus of variations when the side constraints are finite equations. In introducing the rule in the variational calculus [Courant and Hilbert, 1953, 221] write: 'Up to now [i.e. in the case of finite constraints] the multiplier rule has been used merely as an elegant artifice. But multipliers are indispensable if the subsidiary condition takes the general form $G(x, y, z, y', z')=0$, where the expression $G(x, y, z, y', z')$ cannot be obtained by differentiating an expression $H(x, y, z)$ with respect to x, i.e. where G is a *nonintegrable differential expression*'.

In art. 181 of the *Théorie*, Lagrange formulated the multiplier rule for problems of constrained optimization in the calculus of variations. In the *Leçons* he provided an extended treatment of this subject, including the presentation of detailed examples. The multiplier rule proved to be an extremely powerful and effective tool, enabling one to derive results that could only be obtained with considerable difficulty otherwise.

In a variational problem with two dependent variables it is necessary to optimize the primitive or integral of $f(x, y', y'', \ldots, z, z', z'', \ldots)$ evaluated between $x=a$ and $x=b$. The solution must satisfy the Euler–Lagrange differential equations, which Lagrange wrote as

$$f'(y)-\big[f'(y')\big]'+\big[f'(y'')\big]''-\big[f'(y''')\big]'''+\&c=0, \qquad (42)$$

$$f'(z)-\big[f'(z')\big]'+\big[f'(z'')\big]''-\big[f'(z''')\big]'''+\&c=0. \qquad (43)$$

(In modern notation the term $[f'(y)]'$ is $d(\partial f/\partial y')/dx$.) To obtain these equations Lagrange used algebraic, analogical reasoning very different from the modern methods of real analysis.

Suppose now that the variables $x, y, z, \ldots$ satisfy a constraint of the form

$$\phi(x, y, y', \&c., z, z', \&c.) = 0, \tag{44}$$

consisting of a differential equation of arbitrary order connecting the variables of the problem. To obtain the Euler–Lagrange equations in this situation, we multiply (44) by the multiplier function $\Delta(x)$ (note that Δ is a function of x and not a constant) and form the differential equations

$$\begin{aligned} &f'(y) - \left[f'(y')\right]' + \left[f'(y'')\right]'' - \left[f'(y''')\right]''' + \&c + \Delta\phi'(y) \\ &\quad - \left[\Delta\phi'(y')\right]' + \left[\Delta\phi'(y'')\right]'' - \&c. = 0; \end{aligned} \tag{45}$$

$$\begin{aligned} &f'(z) - \left[f'(z')\right]' + \left[f'(z'')\right]'' - \left[f'(z''')\right]''' + \&c + \Delta\phi'(z) \\ &\quad - \left[\Delta\phi'(z')\right]' + \left[\Delta\phi'(z'')\right]'' - \&c. = 0. \end{aligned} \tag{46}$$

In the *Leçons*, Lagrange applied the multiplier rule to two examples involving the motion of a particle descending through a resisting medium. These examples had originally appeared in Chapter 3 of Euler's *Methodus inveniendi* (1744) and concerned the brachistochrone and the curve of maximum terminal velocity (§12.2). Assume the y-axis is measured horizontally and the x-axis is measured vertically downward, and let z equal the square of the speed. We have the dynamical constraint equation

$$z' - 2g + 2\phi(z)\sqrt{1 + y'^2} = 0, \tag{47}$$

giving z as a function of x, y and y'. In each of the examples in question Lagrange wrote down the variational equations (45)–(46). Using them and the constraint equation he obtained differential equations for the multiplier function and the trajectory. Because no restriction was placed on the end value of z in the class of comparison arcs, we obtain another equation, one that allows us to calculate a constant appearing in the expression for the multiplier.

Lagrange was led by means of his multiplier rule to quite straightforward solutions of these problems, arrived at independently of the specialized methods that had appeared in Euler's and his own earlier writings. It is reasonable to assume that the successful treatment of such advanced examples would have confirmed in his mind the validity of the rule and instilled a confidence in the basic correctness of the analytical procedure involved in its application.

In the later calculus of variations the multiplier rule would assume an even more fundamental role and become the basic axiom of the whole subject. Alfred Clebsch (1833–1872) showed how the multiplier rule can be used to reduce problems with higher-order derivatives to problems involving only first derivatives and side constraints [Clebsch, 1858]. In the modern subject, any problem with side constraints is known as a Lagrange problem

and it is solved by means of the multiplier rule. The most general problem of the calculus of variations can be formulated as such a problem and solved in principle using the rule.

6 CALCULUS OF VARIATIONS: SUFFICIENCY RESULTS

In arts. 174–178 Lagrange took up the question of sufficiency in the calculus of variations. Given that a proposed solution satisfies some of the conditions of the problem, it is necessary to investigate what additional conditions must hold in order that there be a genuine maximum or minimum. Here Lagrange reported on results of Adrien-Marie Legendre (1752–1833) published as [Legendre, 1788], and added some important new observations of his own.

For simplicity we consider the case where there is only one dependent variable and where only the first derivative appears in the variational integrand. In a problem in the calculus of variation, a proposed solution will be optimal for Legendre if the sign of the second variation is unchanged (always positive for a minimum, or always negative for a maximum) with respect to all comparison arcs. Let the increment or variation of y be the function $w(x)$. The second variation I_2 is by definition

$$I_2 = \int_{x_0}^{x_1} \left(\frac{\partial^2 f}{\partial y^2} w^2 + 2 \frac{\partial^2 f}{\partial y \partial y'} w w' + \frac{\partial^2 f}{\partial y'^2} w'^2 \right) dx. \tag{48}$$

It is necessary to investigate the sign of I_2. Let $v = v(x)$ be a function of x and consider the expression

$$\frac{d}{dx}(w^2 v), \tag{49}$$

Because $w(x_0) = w(x_1) = 0$ the integral of (49) is zero:

$$\int_{x_0}^{x_1} \frac{d}{dx}(w^2 v)\, dx = 0. \tag{50}$$

We introduce some standard abbreviations for the second partial derivatives:

$$P = \frac{\partial^2 f}{\partial y^2}, \qquad Q = \frac{\partial^2 f}{\partial y \partial y'}, \qquad R = \frac{\partial^2 f}{\partial y'^2}. \tag{51}$$

If we add the integral of (49) to the expression for the second variation I_2 given in (48) there results no change in its value:

$$I_2 = \int_{x_0}^{x_1} \left((P + v') w^2 + 2(Q + v) w w' + R w'^2 \right) dx. \tag{52}$$

The integrand is a quadratic expression in w and w'. Legendre observed that it will become a perfect square if

$$R(P + v') = (Q + v)^2. \tag{53}$$

For $v(x)$ satisfying this differential equation the second variation becomes

$$I_2 = \int_{x_0}^{x_1} R\left(w' + \frac{Q+v}{R}w\right)^2 dx. \tag{54}$$

It is evident that the given transformation is only possible if $R = \partial^2 f/\partial y'^2$ is non-zero on the interval $[x_0, x_1]$. The proposed solution will indeed be a minimum if on the interval we have

$$\frac{\partial^2 f}{\partial y'^2} > 0, \tag{55}$$

which would become known in the later subject as 'Legendre's condition'.

In order to arrive at the expression (54) for I_2 and the associated condition (55) on $\frac{\partial^2 f}{\partial y'^2}$ it is necessary to show that solutions to the differential equation (53) exist and remain finite on the given interval. In his study of the second variation in the *Théorie*, Lagrange called attention to this point and produced examples in which no finite solutions exist (pp. 206–210). Suppose for example that $f(x, y, y') = y'^2 - y^2$. In this case $P = -2$, $Q = 0$ and $R = 2$ and (54) becomes $2(v' - 2) = v^2$. By elementary methods this equation may be integrated to produce $v = 2\tan(x + c)$, where c is a constant. It is clear that if $x_1 - x_0$ is greater than $\pi/2$, then no solution of (53) will exist.

Lagrange's exposition of Legendre's theory was important because it made known to a wide audience results that likely would otherwise have remained buried in the memoirs of the Paris *Académie*. Carl Gustav Jacobi (1804–1851), in his ground-breaking paper [1837] on sufficiency theory, began his investigation with Lagrange's formulation of the subject. Lagrange's discussion of the solutions of (53) also raised new considerations that were important stimuli for Jacobi's investigation.

7 CONCLUSION

An important and under-appreciated contribution of the history of mathematics is to provide insight into the foundations of a mathematical theory by identifying the characteristics of historically earlier formulations of the theory. The conceptual relativism of scientific theories over history is one of the major findings of the history of science since Thomas Kuhn. In this respect, history of mathematics possesses a special interest lacking in the history of other branches of science, because earlier mathematical theories are enduring objects of technical interest and even of further development.

What is primarily of note in Lagrange's *Théorie*, beyond its substantial positive achievements, is that it developed analysis from a perspective that is different from the modern one. It did so in a detailed and sophisticated way, with a self-conscious emphasis on the importance of building a sound foundation. A product of the intellectual milieu of advanced research at the end of the 18th century, it stands at the cusp between algebraic and modern analysis. It retains an historical interest for us today that transcends its contribution to technical mathematics.

Although the foundation Lagrange proposed did not achieve final acceptance, his conception of analysis exerted considerable influence in the 19th century. Cauchy was able to adapt many of Lagrange's results and methods in developing an arithmetical basis for the calculus [Grabiner, 1981]—while also refuting his belief in the generality of (1) (§25). The spread of Continental analysis to Britain relied heavily on Lagrange's writings. The formal algebraists George Peacock in England and Martin Ohm in Germany were influenced by his mathematical philosophy. Lagrange's insistence that analysis should avoid geometrical and mechanical ideas was taken up with some emphasis by the Bohemian philosopher Bernhard Bolzano. The school of researchers who used operator methods in the theory of differential equations, François Servois in France, and George Boole and Duncan Gregory in Britain, took inspiration from Lagrange's writings on analysis (§36.2). Finally, even after the consolidation of Cauchy's arithmetical foundation, Lagrange's emphasis on the algorithmic, operational character of the calculus continued to inform writers of textbooks as well as non-mathematicians such as engineers and physicists who used calculus in their research and teaching.

BIBLIOGRAPHY

Arbogast, L. 1800. *Du calcul des dérivations et de ses usages dans la théorie des suites et dans le calcul différentiel*. Strasbourg: Levrault.

Clebsch, R.F.A. 1858. 'Ueber die Reduktion der zweiten Variation auf ihre einfachste Form', *Journal für die reine und angewandte Mathematik, 60*, 254–273.

Courant, R. 1937. *Differential and integral calculus*, vol. 1, 2nd ed., New York: Interscience Publishers. [Trans. of German original.]

Courant, R. and Hilbert, D. 1953. *Methods of mathematical physics*, vol. 1, New York: Interscience Publishers. [Trans. of German original.]

Fraser, C. 1981. 'The approach of Jean D'Alembert and Lazare Carnot to the theory of a constrained dynamical system', Ph.D. dissertation, University of Toronto.

Fraser, C. 1985. 'J.L. Lagrange's changing approach to the foundations of the calculus of variations', *Archive for the history of exact sciences, 32*, 151–191.

Fraser, C. 1987. 'Joseph Louis Lagrange's algebraic vision of the calculus', *Historia mathematica, 14*, 38–53.

Grabiner, J. 1981. *The origins of Cauchy's rigorous calculus*, Cambridge, MA: MIT Press.

Jacobi, C.G.J. 1837. 'Zur Theorie der Variations-Rechnung und der Differential-Gleichungen', *Journal für die reine und angewandte Mathematik, 17*, 68–82.

Lagrange, J.L. *Works*. *Oeuvres*, 14 vols., Paris: Gauthier–Villars, 1867–1892. [Repr. Hildesheim: Olms, 1968.]

Lagrange, J.L. 1774. 'Sur une nouvelle espèce de calcul relatif à la différentiation & à l'intégration des quantités variables', *Nouveaux mémoires de l'Académie des Sciences et Belles-Lettres* (*Berlin*) (1772), 185–221. [Repr. in *Works*, vol. 3, 441–476.]

Lagrange, J.L. 1788. *Méchanique analitique*, 1st ed., Paris: Desaint. [2nd ed., 'Mécanique analytique', 1811–1815, 1853–1855, and *Works*, vols. 11–12 (§16).]

Lagrange, J.L. 1798. *Traité de la résolution des équations numériques de tous les degrés*, 1st ed., Paris: Courcier. [2nd ed. 1808. 3rd ed. 1826; also *Works*, vol. 8.]

Lagrange, J.L. 1801. 'Leçons sur le calcul des fonctions', *Séances des Ecoles Normales, 10*, 1–534. [Repr. as *Journal de l'Ecole Polytechnique, 5, cahier* 12 (1804).]

Lagrange, J.L. 1806. *Leçons sur le calcul des fonctions*, new ed., Paris: Courcier. [Repr. as *Works*, vol. 10. Includes 'a complete treatise on the calculus of variations'.]

Legendre, A.-M. 1788. 'Sur la manière de distinguer les maxima des minima dans le calcul des variations', *Mémoires de l'Académie des Sciences de Paris*, (1786), 7–37.

Ovaert, J.-L. 1976. 'La thèse de Lagrange et la transformation de l'analyse', in C. Houzel *et alii*, *Philosophie et le calcul de l'infini*, Paris: Maspero, 159–222.

Poinsot, L. 1806. 'Théorie générale de l'équilibre et du mouvement des systèmes', *Journal de l'Ecole Polytechnique, 6, cahier 13*, 206–241. [Critical ed. P. Bailhache: Paris: Vrin, 1975.]

CHAPTER 20

S.F. LACROIX, *TRAITÉ DU CALCUL DIFFÉRENTIEL ET DU CALCUL INTÉGRAL*, FIRST EDITION (1797–1800)

João Caramalho Domingues

In this encyclopaedic work Lacroix compiled and organized much of the knowledge of the time on the differential, integral, and finite difference calculi.

First publication. *Traité du calcul différentiel et du calcul intégral* and *Traité des différences et des séries*, 2 + 1 vols. Paris: J.M. Duprat, 1797, 1798, 1800. xxxii + 524; viii + 732; viii + 582 pages. [Available at *Gallica*: http://gallica.bnf.fr.]

Second edition. *Traité du calcul différentiel et du calcul intégral*, 3 vols. Paris: V.^{e} Courcier, 1810, 1814, 1819. lvi + 653; xxii + 820; xxiv + 776 pages.

German translation of volumes 1 and 2 of the 1st ed. Lehrbegriff des Differential- und Integralcalculs (trans. J.P. Grüson). 2 vols. Berlin: F.T. Lagarde, 1799, 1800.

Related articles: Euler on analysis and the calculus (§12, §13), Lagrange on the calculus (§19), Cauchy on real-variable analysis (§25).

1 MATHEMATICAL TEACHER AND WRITER

Sylvestre François Lacroix was born in Paris in 1765, to a modest family. His first teacher of mathematics was the abbé Marie, but his great educational influence seems to have been Gaspard Monge, whose free courses he attended at least in 1780. It was Monge (1746–1818) who secured in 1782 Lacroix's first teaching post: professor of mathematics at the *École des Gardes de la Marine* in Rochefort. He stayed in Rochefort until 1785. His first attempts at research, from 1779, had consisted of long astronomical calculations, but in Rochefort, although not abandoning astronomy entirely, he studied partial differential equations and their application to surface theory. These studies were done under Monge's supervision, through continued correspondence.

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

Lacroix was not happy far from Paris and Monge convinced the Marquis de Condorcet (1743–1794) to employ Lacroix as his substitute at the newly founded *Lycée* (not to be confused with the subsequent secondary education institutions; this *Lycée* was a private school for gentlemen who wished to acquire a general culture; it had renowned professors who in fact nominated their substitutes to give all lectures under their direction; Condorcet was in charge of mathematics). Condorcet became a second great influence for Lacroix. Together they prepared a new edition of Euler's *Lettres à une princesse d'Allemagne*. Lacroix started teaching at the *Lycée* in January 1786. In February 1787 he accumulated that with teaching at the *École Royale Militaire de Paris*. It was then that he started gathering material for the *Traité*.

In 1787 the mathematics course at the *Lycée* was abolished due to lack of students and in 1788 the *École Militaire* was closed. Lacroix was forced to go once again into *exile*, taking an appointment at the *École Royale d'Artillerie* in Besançon. From there he maintained postal contact with Monge, Condorcet, J.D. Cassini, J.J. Lalande, A.M. Legendre and P.S. Laplace. In 1789 he was elected Condorcet's correspondent by the *Académie des Sciences*. Lacroix kept gathering material for his planned *Traité*, but in 1792 he complained in a letter to Laplace about the scientific indigence of Besançon. He also asked Laplace to send him offprints of some of his memoirs.

Lacroix returned definitively to Paris in 1793, succeeding Laplace as *examinateur* of candidates and students at the *Corps d'Artillerie*. He was then able to complete the *Traité*. Printing started in 1795, although the 1st volume only appeared in 1797. Until his death in 1843 Lacroix followed a brilliant educational career: in 1794 he belonged to the *Commission de l'Instruction Publique*, and afterwards he held examiner and/or teaching posts at the *École Polytechnique*, *École Normale* (*de l'an III*), *École Centrale des Quatre Nations*, *Faculté des Sciences de Paris* (of which he was also dean), and finally at the *Collège de France*.

A consequence of his teaching at the *École Centrale des Quatre Nations* (a secondary school) was his writing a series of seven textbooks, from a *Traité élémentaire d'arithmétique* to a *Traité élémentaire du calcul différentiel et du calcul intégral* (not to be confused with the non-*élémentaire* one that is the subject of this article), although the last of these was written to be used elsewhere: at the *École Polytechnique*, where Lacroix also taught. These textbooks were highly successful throughout the 19th century, reaching impressive numbers of editions (the *Arithmétique* reached the 20th in 1848; the *Éléments d'algèbre* and the *Éléments de géométrie* both reached the 19th, in 1849 and 1874, respectively) and being translated into several languages.

In spite of an insignificant research career, Lacroix was respected in the mathematical community, being elected in 1799 as a member of the first class of the *Institut National* (which had replaced the *Académie des Sciences*). According to Taton, when Lacroix started writing his large treatises and his textbooks, '[h]e understood that his so wide erudition and his so remarkable talent for clarification [*mise au point*] and presentation would allow him to make there a work more useful than that he would have achieved had he confined himself to researches on details' [Taton, 1953a, 590].

2 PURPOSES OF THE *TRAITÉ*

As has already been said, Lacroix started gathering material for his treatise in 1787, while employed at the *École Royale Militaire de Paris*. Apparently the reason for which he felt this treatise would be useful, was the enormous gap between the elementary books available and the research memoirs on calculus subjects. In other words, there was no advanced (say, modern-day graduate level) textbook on the calculus. The popular works like the one by Étienne Bézout were too elementary; and the only advanced comprehensive survey (Euler's set of works) was getting outdated. It took much time and effort for a young man living in Paris and wishing to pursue mathematics to read all the necessary works to bridge the gap between elementary calculus and research-level calculus: those works were memoirs dispersed in academic journals and books with low print runs. And for someone living outside Paris, this would be nearly impossible, due to the lack of good libraries (1st ed., vol. 1, iii; 2nd ed., vol. 1, xviii–xix).

Another stimulus was Lacroix's reading of Lagrange's memoir [1774] 'Sur une nouvelle espèce de calcul relatif à la differentiation et à la intégration des quantités variables', where a suggestion was made for a new foundation of the calculus. Lacroix envisioned then a grand plan: not only to compile all the major methods, but also to choose between different but equivalent ones or to show how they relate to one another, as well as to give all of them a uniform hue that would not allow to trace the respective authors. It was clearly intended to replace Euler's set of works, but not to be a first introduction to the calculus: 'such a voluminous treatise as this one, can hardly be consulted but by people to whom the subject is not entirely new, or that have an unwavering taste for this kind of study' (2nd ed., vol. 1, xx). The result was a monumental reference work: the first edition has around 1800 pages in total; an *encyclopédiste* appraisal of the calculus at the turn of the century.

Two features that seem to be unprecedented in mathematical books should be pointed out. One is the remarkable bibliography attached to the table of contents: for each chapter and section, Lacroix gives a list of the main works related to its subject. All the major 18th-century works on the calculus are included there, as well as many minor and even some obscure ones. Typically, in the list for a given chapter/section one will find the corresponding chapters in one of Euler's three books, some other relevant books (say, Lagrange's *Théorie des fonctions analytiques*, Jacob Bernoulli's *Opera* or Stirling's *3rd-order lines*) and memoirs drawn from the volumes published by the *Académie des Sciences de Paris*, by the Berlin Academy, by the St. Petersburg Academy, by the Turin Academy, and so on. The most cited authors seem to be those that one expects: Euler, Lagrange, Laplace, J. Le R. d'Alembert, Monge; but it is also possible to find references to such authors as Fagnano (1st ed., vol. 2, v) or Oechlitius (1st ed., vol. 3, viii). As an example, Chapter 2 of the second volume, which is dedicated to the calculation of areas, volumes and arc-lengths, has a bibliography with 20 authors, and about 35 works (excluding some historical references).

Another remarkable feature is the subject index included at the end of the third volume. Not only it may well be the first subject index in a mathematical book, but it is 34 pages long (1st ed., vol. 3, 545–578)!

3 DIFFERENTIAL CALCULUS

Table 1 summarizes the contents of the first volume, dedicated to the differential calculus. It also compares the first with the second edition, but for now we will only concern ourselves with the first one.

As has already been said, printing of the first volume started in 1795. It was interrupted for some time and only concluded in 1797. According to Lacroix (p. xxiv) this meant that

Table 1. Volume I of Lacroix's treatise.

1st edition		2nd edition		(Some) topics covered
Chapter	Pages	Chapter	Pages	
Preface.	iv–xxix	Preface.	i–xlviii	History of the calculus.
Table of contents.	xxx–xxxii	Table of contents.	xlix–lvi	Contents and bibliography.
Introduction.	1–80	Introduction.	1–138	Functions, series and limits; series expansion of functions.
Ch. 1: Principles of differential calculus.	81–194	Ch. 1: Principles of differential calculus.	139–248	Differentiation of functions; differentiation of equations.
Ch. 2: Analytic uses of differential calculus.	195–276	Ch. 2: Use of diff. calculus to expand functions.	249–326	Differential methods for expansion of functions in series.
		Ch. 3: Particular values of diff. coefficients.	327–388	Indeterminacies; maxima and minima.
Ch. 3: Algebraic equations.	277–326			Symmetric functions; complex numbers.
Ch. 4: Curve theory.	327–434	Ch. 4: Curve theory.	389–500	Analytic and differential geometry of plane curves.
Ch. 5: Curved surfaces and curves of double curvature.	435–519	Ch. 5: Curved surfaces and curves of double curvature.	501–652	Analytic and differential geometry of surfaces and space curves.

he profited little from the lectures Lagrange gave on the same subject at the *École Polytechnique* in 1795–1796 and which gave origin to Lagrange's *Théorie des fonctions analytiques* (§19), also published in 1797. It seems therefore that Lacroix had to develop the suggestion of [Lagrange, 1774] in a mostly independent way. But there is some evidence that the little profit was not null.

The first volume starts with a Preface, which includes an explanation of the aims of the work and the plan for the three volumes, but is mostly taken by a long account of the history of the calculus (pp. iv–xxiii). After the table of contents (with bibliography) comes an Introduction. Its purpose is to give 'series expansions of algebraic, exponential, logarithmic and trigonometric functions' by algebraic means, without recourse to the notion of infinity (p. xxiv). It is clearly intended to be equivalent to the first volume of Euler's *Introductio* (§13). It also corresponds, with Chapter 3, to what would be known for some time, at least in the curriculum of the *École Polytechnique*, as *algebraic analysis* (§25.2). Lacroix describes it as 'the intermediary analysis between the elements of algebra proper, and the differential calculus' (2nd ed., vol. 1, xx).

Lacroix starts by defining *function*: 'Any quantity the value of which depends on one or more other quantities is said to be a *function* of these latter, whether or not it is known which operations are necessary to go from them to the former' (p. 1). But the example given for a function in which the necessary operations are not known is the root of a fifth degree equation. In fact this definition means that a function had to be defined explicitly or implicitly using the usual mathematical operations. As in Euler (§14.2), it is assumed that any function has a power-series expansion. The Eulerian classification of functions as explicit or implicit, algebraic or transcendental is also followed.

Although there is some discussion on limits and convergence of series (pp. 4–18), most of the Introduction is concerned with power-series expansions of the most common functions. For this a 'weak' version of the binomial theorem, stating

$$(1+x)^n = 1 + nx^{n-1} + \text{etc.} \tag{1}$$

is proven (for 'any n'; the full expansion is given for integer n). It is widely used, along with the method of indeterminate coefficients.

In Chapter 1 Lacroix builds the differential calculus on the basis suggested in [Lagrange, 1774]. First comes the expansion

$$f(x+k) - f(k) = X_1 k + X_2 k^2 + X_3 k^3 + \text{etc.} \tag{2}$$

Then, after establishing the iterative relation between the coefficients and thus renaming them to

$$f(x+k) - f(k) = f'(x)k + \frac{f''(x)}{2}k^2 + \frac{f'''(x)}{1\cdot 2\cdot 3}k^3 + \text{etc.} \tag{3}$$

the first term $f'(x)k$ is christened *differential* 'because it is only a portion of the difference' and is given the symbol $df(x)$. 'For uniformity of symbols [...] dx will be written instead of k', so that

$$f'(x) = \frac{df(x)}{dx} \tag{4}$$

is an immediate *conclusion*.

Sometimes $f'(x)$, $f''(x)$, etc. are called 'derived functions', because of the derivation process that relates each of them to the previous, but the name they gain on p. 98 (and which will be used throughout the three volumes) is *differential coefficients*. The differential notation will also be much more frequent. Overall this foundation for the calculus is Lagrangian, but much closer to [Lagrange, 1774] than to the *Théorie des fonctions analytiques* (§19), where differentials have no place.

The results obtained in the Introduction allow easy deductions of the differentials of one-variable algebraic, logarithmic, exponential and trigonometric functions: it is only necessary to expand $f(x+dx)$ and extract the term with the first power of dx.

Differentiation of functions of two variables is also inspired by [Lagrange, 1774], but without resorting to the cumbersome notation Lagrange had employed ($u'^{,''}$ for our $\frac{\partial u^3}{\partial x\,\partial y^2}$). $f(x+h, y+k)$ is expanded in two steps and in two ways (via $f(x+h, y)$ and via $f(x, y+k)$), whence the conclusion that $\frac{d^2u}{dx\,dy} = \frac{d^2u}{dy\,dx}$. The definition of differential as the first-order term in the expanded series of the incremented function is extended to $u = f(x, y)$ giving

$$df(x, y) = du = \frac{du}{dx}\,dx + \frac{du}{dy}\,dy \tag{5}$$

(the ∂ notation is still absent, but proper warning is given about the fact that $\frac{du}{dx}\,dx$ is the differential of u regarding only x as variable and not to be confused with du).

Lacroix occupies a considerable amount of space (pp. 134–189) with differentiation of equations. As in Euler (§14), this is both a manner of dealing with implicit functions and of preparing the way for the treatment of differential equations in the integral calculus: for instance, condition equations for exact differentials are first handled here. Lacroix himself acknowledges having borrowed much from Euler in this chapter (p. xxiv), and that influence is clear in its structure and in all that does not relate directly to foundations.

Chapter 1 ends with a section about alternative foundations for the calculus. Both d'Alembert's limit approach and Leibniz's infinitesimals are treated. This is typical of Lacroix's *encyclopédiste* approach: to expound all relevant alternative methods or theories. It is also an essential instance of that approach because in future chapters Lacroix will sometimes need to resort to one or other of those alternative foundations in order to explain some particular method.

Chapter 2 is dedicated to some analytic questions around the differential calculus. First comes its employment in expanding functions in series, for which of course Taylor's theorem is central. After this comes an examination of certain cases in which the differential coefficient 'becomes infinite' (as with $f(x) = \sqrt{x-a}$ for $x = a$) and why the expansion (2), 'although true in general', is not valid in such cases. The explanation for this rests on the irrationality of the function involved disappearing for certain values of the variable, dragging a collapse of multiple values of the function. Lacroix attributes this to Lagrange and in fact it appears in his *Théorie des fonctions analytiques*: it may be one of the few remarks drawn from Lagrange's lectures at the *École Polytechnique* that Lacroix

was able to include in the first volume. This is followed by a discussion of indeterminacies ($\frac{0}{0}$, $0 \times \infty$, ...) and how to raise them; and by another on maxima and minima of functions of one or several variables.

Chapter 3 is an algebraic interlude, on certain matters relating to equations that were absent from elementary books: a section on symmetric functions of the roots of an equation, which would later migrate to Lacroix's *Complément des élémens d'algèbre* (1st ed., 1800) and was therefore removed in the second edition of the *Traité*; and another on complex numbers ('imaginary expressions'), including the fundamental theorem of algebra, which would move to the Introduction in the second edition.

The two final chapters are devoted to analytic and differential geometry: Chapter 4 on the plane; Chapter 5 in the space. Here the influence from Monge is most marked.

What was still generally known as 'application of algebra to geometry' was then being transformed into *analytic geometry*. Monge was the main architect of this change (with an important suggestion by Lagrange in a 1773 memoir on tetrahedra), but Lacroix played an important role in its systematization, precisely in this *Traité* [Taton, 1951, ch. 3]). As he explains in the Preface, he tried to keep apart all geometric constructions and synthetic reasonings, and to deduce all geometry by purely analytic methods. That is why Chapter 4 starts by an extensive study of fundamental formulae for points, straight lines and distances, to be used in what follows, instead of 'geometric constructions'. These elementary subjects were usually regarded as belonging to the realm of synthetic geometry. By the second edition, Lacroix had published an elementary book that included such matters (*Traité de Trigonométrie* [...] *et d'application de l'Algèbre à la Géométrie*) and was therefore able to suppress this section.

After those preliminaries Lacroix develops the analytic geometry of plane curves, including plotting, classification of singular points and changes of coordinates. Changes of coordinates have several applications, including finding tangents and multiple points. Before differential geometry properly speaking, comes the application of series expansions (which because of their approximative nature supply a way of finding tangents and asymptotes).

The central part of Chapter 4 is the application of differential calculus (that is, the use of differential coefficients) to find properties of the curves: their tangents, normals, singular points, the differentials of their arc-length and of the area under them; and to develop a theory of osculation and hence of curvature via the osculating circle.

The chapter concludes in a manner very typical of Lacroix: presenting alternative points of view, namely an application of the method of limits to find tangents and osculating lines and the Leibnizian consideration of curves as polygons. It is significant that in total this chapter has five approaches to the search of tangents. In this last section is included a study of envelopes of one-parameter families of curves, the language alternating between limit-oriented and infinitesimal. A very important special case is that of the evolute of a given curve, formed by the consecutive intersections of its normals.

The matter of Chapter 5, a theory of surfaces and space curves, is mostly due to Monge, according to Lacroix (p. 435). The fundamental formulae for planes and points, straight lines and distances in space are followed by a discussion on understanding the shape of second-order surfaces (namely by plane cuts), and by changes of coordinates.

There is some discussion of contact of surfaces using their series expansions, but the differential study of a surface is done mainly by taking plane sections through the point one is considering. Alternatively to comparison of coefficients in series expansions, the tangent plane through a point with coordinates x', y', z' is determined by the tangents to the sections parallel to the vertical coordinate planes (these tangents have slopes $\frac{dz'}{dx'}$, $\frac{dz'}{dy'}$, so that

$$z - z' = \frac{dz'}{dx'}(x - x') + \frac{dz'}{dy'}(y - y') \tag{6}$$

is the equation of the plane). Osculating spheres are studied similarly.

Not surprisingly, curvature of a surface on a point is studied through the radii of curvature of plane sections through that point: these have a maximum and a minimum, which allow to calculate the curvature of any other plane section. There is no discussion yet of kinds of curvature or of the possibilities of the centers of curvature being on the same or on different sides of the surface.

Envelopes of one-parameter families of surfaces are studied as the 'limits' of their consecutive intersections (these intersections are called, following Monge, 'characteristics'). A special case is that in which the generating surfaces are planes: the envelope is then called a 'developable surface'.

Three approaches are given to study curves in space ('curves of double curvature'). But two of them only briefly (through their projections on the coordinate planes; and through the series expansions of two coordinates as functions of the third). The bulk of the section follows Monge in regarding space curves as polygons where three consecutive sides are not coplanar. This allows not only the study of tangents, osculating planes, and differentials of arc-length, but also of the developable surface generated by a curve's normal planes, and of evolutes.

4 INTEGRAL CALCULUS

Although the second volume of Lacroix's *Traité* (Table 2) is the largest of the three, it is the one that gets less attention from Lacroix in the general Preface at the beginning of volume I. The integral calculus, being just the inverse of the differential calculus, did not offer much occasion for reflection: it consisted only of a 'collection of analytical procedures, which is enough to order so as to make perceive their connections' (1st ed., vol. 1, xxvii). Lacroix proposes then to follow Euler's ordering in his *Institutionum calculi integralis* (1768–1770), adding new developments and replacing some methods by more recent ones.

Most of Chapter 1 is dedicated to find antiderivatives of functions of one variable: algebraic, rational, irrational, and transcendental (exponential, logarithmic and trigonometric). There is also some attention to approximation techniques. Term-by-term integration of power series is both a way of finding exact antiderivatives and of approximating them. But since the resulting series is not always convergent, a 'general method' (taken from Euler's *Institutionum*) is given to approximate integrals by *rectangles*:

Table 2. Volume II of Lacroix's treatise.

1st edition		2nd edition		(Some) topics covered
Chapter	Pages	Chapter	Pages	
Table of contents.	iii–viii	Table of contents.	vii–xxi	Contents and bibliography.
Ch. 1: Integration of functions of one variable.	1–160	Ch. 1: Integration of functions of one variable.	1–155	Antiderivatives; approximate values.
Ch. 2: Quadratures, cubatures and rectifications.	161–220	Ch. 2: Quadratures, cubatures and rectifications.	156–224	Areas, volumes and arc-lengths.
	(partly in chs. 3 and 4)	Ch. 3: Integration of diff. functions of several variables.	225–249	Conditions for integrability.
Ch. 3: Integration of differential equations in two variables.	221–452	Ch. 4: Integration of diff. eqs. on two variables.	250–372	Solutions of ordinary differential equations.
		Ch. 5: Particular solutions of diff. eqs.	373–408	Singular solutions.
		Ch. 6: Approximate integr. of diff. eqs.	409–446	Approximation methods.
		Ch. 7: Geometric applications of diff. eqs. on two vars.	447–470	Geometrical problems.
		Ch. 8: Comparison of transcendental functions.	471–502	Logarithmic, trigonometric and elliptic functions.
Ch. 4: Integration of functions of two or more variables.	453–654	Ch. 9: Integration of equations on three or more variables.	503–720	Partial differential equations.
Ch. 5: Method of variations.	655–724	Ch. 10: Method of variations.	721–816	Calculus of variations.

$$Y + f(a)(a_1 - a) + f(a_1)(a_2 - a_1) + \cdots + f(a_{n-1})(a_n - a_{n-1}) \tag{7}$$

(where Y is the value of $\int f(x)\,dx$ at $x = a$, and the differences $a_i - a_{i-1}$ are not necessarily all equal).

In this connection a distinction appears between *indefinite* and *definite integrals*, along with the respective definitions. According to [Cajori, 1919, 272] this was the first time such definitions were given. However, Lacroix himself attributes them, rather vaguely, to 'the Analysts' (p. 142), presumably [Laplace, 1782].

Chapter 2, dedicated to areas, arc-lengths and volumes, consists mainly of examples, since the differentials have already been found in the first volume and the methods of integration have been studied in Chapter 1. But it ends with a small section on squarable curves (that is, functions with algebraic integrals).

Chapter 3 is one of the central parts of the volume: ordinary differential equations. Naturally separation of variables and the use of integrating factors are the main methods of solution for first-order equations. For second-order equations, integrating factors are also used, alongside with certain reductions to first order when possible. Linear equations (and systems of linear equations) get some attention, but Lacroix rejects that name and prefers to call them just 'first-degree equations', since the word linear refers to straight lines, to which these equations do not relate.

Singular solutions are examined, following Lagrange's explanation of 1774 (but using Laplace's term: 'particular solutions'; Lagrange had called them 'particular integrals'), and including their geometrical interpretation as envelopes of the families of curves given by the 'complete integral'. There also some methods to approximate solutions of first- and second-order equations, including one that gives them in the form of continued fractions.

Here occurs a very interesting remark, although also very casual. After giving

$$Y + f'(a)(a_1 - a) + f'(a_1)(a_2 - a_1) + \cdots + f'(a_{n-1})(a_n - a_{n-1}) \tag{8}$$

as an approximate solution to a first-order differential equation (analogously to (7), $f'(a_i) = \frac{dy}{dx}|_{x=a_i}$ being taken from the equation), Lacroix notes that this means that every first-order equation is possible, that is, that it is possible to assign a solution, 'either rigorous, or approximate' (p. 287). This looks like a very crude attempt at an existence theorem.

This chapter also includes some geometrical applications (orthogonal trajectories, for instance), and one analytical application: the study of transcendental functions (logarithmic, trigonometric and elliptic) starting from their differential equations.

The other major chapter in this volume is Chapter 4, on the 'integration of functions of several variables'. It starts with a section on 'total differential equations': those that involve all the differential coefficients of some order (in other words, all the partial derivatives of some order) of the function we seek (either explicitly or implicitly). But almost all of the chapter is devoted to those that do not: 'partial differential equations'.

The types of partial differential equations studied are: first order and first degree (Lacroix still rejecting the word 'linear'); first order and degree greater than one (by reduction to first degree); first degree and any order; and a few cases of second order and degree greater than one. Singular (i.e., 'particular') solutions are once again studied following Lagrange, as Lagrangian is a distinction between 'complete integrals' (with as

many constants as the order of the equation) and 'general integrals' (involving arbitrary functions).

The last part of the chapter draws inspiration from Monge. It includes the 'geometrical construction' of a partial differential equation on three variables (that is, a study of the surfaces that obey that equation); and a study of 'total differential equations' on three variables that cannot be integrated (to none of the variables can be assigned a function of the other two), which result in families of curves that do not form a surface.

The second volume concludes with a chapter on the calculus of variations. Isaac Todhunter, in his *History of calculus of variations*, summarizes a review of this chapter by commenting that this subject 'does not seem to have been very successfully expounded by Lacroix, and this is perhaps one of the least satisfactory parts of his great work' [Todhunter, 1861, 27]. He also quotes a complaint from another author (a Mr Abbatt) that on this subject Lacroix was 'prolix and inelegant'. Todhunter only saw the second edition of Lacroix's *Traité*, but I think his criticism might apply to the first edition as well.

In this first edition, Lacroix makes no attempt to suit the calculus of variations to the Lagrangian power-series foundation of the calculus, so he presents Lagrange's δ-algorithm in its Leibnizian shape (the rules of δ-differentiation come from those of d-differentiation by plain analogy and $\delta\, dy = d\delta y$ is justified using infinitesimal considerations).

5 DIFFERENCES AND SERIES

At first glance, the status of the third volume as part of the general work is doubtful: it has a different title (*Treatise on differences and series*) and it is a continuation ('faisant suite') of the *Traité du calcul différentiel et du calcul intégral*. But the numbering of its paragraphs follows directly that of the second volume, and the subject index at the end is for the entire set of three volumes.

At the start, Lacroix reminds the reader that in the first two volumes series only occurred as expansions of functions, and their only purposes were to help study certain properties of those functions or else to give approximate values of them. In this volume, series are to be studied for themselves.

It must be noted that Lacroix keeps the 18th-century tradition of not distinguishing between *series* and *sequences*. Both words are used interchangeably.

Chapter 1, occupying more than half the volume, is roughly a discrete version of the differential and integral calculus. It starts by the basic definitions: given a sequence $u, u_1, u_2, u_3, \ldots$,

$$\begin{aligned} &\Delta u = u_1 - u, \qquad \Delta u_{n-1} = u_n - u_{n-1}, \\ &\Delta^m u_{n-1} = \Delta^{m-1} u_n - \Delta^{m-1} u_{n-1}. \end{aligned} \tag{9}$$

Some calculations follow, giving

$$u_n = u + \frac{n}{1}\Delta u + \frac{n(n-1)}{1\cdot 2}\Delta^2 u + \frac{n(n-1)(n-2)}{1\cdot 2\cdot 3}\Delta^3 u + \text{etc.} \tag{10}$$

and

Table 3. Volume III of Lacroix's treatise.

1st edition		2nd edition		(Some) topics covered
Chapter	Pages	Chapter	Pages	
Table of contents.	iii–viii	Table of contents.	vii–xxiv	Contents and bibliography.
Ch. 1: Calculus of differences.	1–300	Ch. 1: Direct calculus of differences.	1–74	Finite differences; interpolation.
		Ch. 2: Inverse calculus of differences of explicit functions.	75–194	Σ-integration; summation of series; interpolation.
		Ch. 3: Integration of difference equations.	195–321	Difference equations.
Ch. 2: Theory of sequences drawn from generating functions.	301–355	Ch. 4: Theory of sequences drawn from generating functions.	322–373	Generating functions.
Ch. 3: Application of integral calculus to the theory of sequences.	356–529	Ch. 5: Application of integral calculus to the theory of sequences.	374–411	Summation of series; interpolation.
		Ch. 6: Evaluation of definite integrals.	412–528	Use of series and infinite products.
		Ch. 7: Definite integrals applied to the solution of differential and difference eqs.	529–574	Transcendental functions.
Ch. 4: Mixed difference equations.	530–544	Ch. 8: Mixed difference equations.	575–600	Difference-differential equations.
Subject index.	545–578	Subject index.	733–771	
Corrections and additions.	579–582	Corrections and additions.	601–732	Extra-typographical corrections.

$$\Delta^n u = u_n - \frac{n}{1}u_{n-1} + \frac{n(n-1)}{1\cdot 2}u_{n-2} - \frac{n(n-1)(n-2)}{1\cdot 2\cdot 3}u_{n-3} + \text{etc.} \tag{11}$$

These results suggest writing the expressions

$$u_n = (1+\Delta u)^n \quad \text{and} \quad \Delta^n u = (u-1)^n \tag{12}$$

'as long as one remembers to change, in the expansion of the first [expression], the exponents of the powers of Δu into exponents of the characteristic Δ, and in the second, to transform the exponents of u into indices'. (10) and (11) were of course quite standard, but their counterparts in (12) come from [Lagrange, 1774], the same memoir where he suggested the power-series foundation for the calculus. Lacroix proceeds following Lagrange on this formalistic path, thus writing, for example

$$\Delta^n u = (e^{\frac{du}{dx}h} - 1)^n \tag{13}$$

(where u is a function of x and $h = \Delta x$), with similar provisions for the exponent. However, while Lagrange had established this only on an analogy basis, Lacroix includes a proof by Laplace. After this comes a section on interpolation, where formulae derived from these are applied, alongside with others, mostly polynomial (e.g., Lagrange interpolating polynomial).

The inverse calculus of differences is presented as a discrete equivalent to the integral calculus: if $\Delta u = f(x, \Delta x)$ then $\Sigma f(x, \Delta x) = u$ is the *integral* of $f(x, \Delta x)$. Σ-integration is developed as far as possible, using among other tools Bernoulli numbers and 'second-order powers': $[p]^n = p(p-1)\cdots(p-n+1)$. Σ-integrals are related to $\int$-integrals in a similar way to what had been done between differences and differentials:

$$\Sigma^m u = \frac{1}{(e^{\frac{du}{dx}h} - 1)^n} \tag{14}$$

(as long as $\frac{du^p}{dx^p}$ is changed into $\frac{d^p u}{dx^p}$ and $\frac{du^{-p}}{dx^{-p}}$ into $\int^p u\, dx^p$).

Results like (10) relate Σ-integration to summation of series. The general relation is

$$Sf(x, \Delta x) = \Sigma f(x, \Delta x) + f(x, \Delta x) - \text{const.} \tag{15}$$

($Sf(x, \Delta x)$ is the sum; the arbitrary constant introduced by integration must now be removed).

A large part of the chapter is dedicated to difference equations, including known methods of solution but also reflections on the arbitrary quantities introduced by them (which need not be constant, unlike those in differential equations) and singular solutions.

The much shorter Chapter 2 is yet another example of the encyclopedic character of this *Traité*: much of the matter of chapter 1 is readdressed, this time using an approach, given in [Laplace, 1782] in the *Mémoires de l'Académie des Sciences de Paris*, based on *generating functions*. u is the generating function of y_x if

$$u = y_0 + y_1 t + y_2 t^2 + \cdots + y_x t^x + y_{x+1} t^{x+1} + \text{etc.} \tag{16}$$

In Chapter 3 integral calculus is applied to series and sequences (namely for summing series and for interpolation), and vice versa (series are used to evaluate definite integrals). Definite integrals get much attention here, which motivates some use of the (Eulerian) notation

$$\int \frac{x^{m-1}\, dx}{1 + x^n} \left[\begin{array}{l} x = 0 \\ x = \text{inf} \end{array}\right] \tag{17}$$

(that is, the integral taken from 0 to $+\infty$). Definite integrals are also used to study some transcendental functions.

Volume 3, and the *Traité*, conclude with a small chapter on 'mixed difference equations', that is, difference-differential equations: an analytical theory is followed by some geometrical applications.

6 THE *TRAITÉ ÉLÉMENTAIRE* AND THE SECOND EDITION OF THE *TRAITÉ*

In 1802 Lacroix, then a *professeur* at the *École Polytechnique*, published a *Traité élémentaire du calcul différentiel et du calcul intégral* [Lacroix, 1802]. According to the publisher's list of works by Lacroix, it was 'partly taken from' the large *Traité*. It does seem to be mostly an abridged version of the latter. It is divided into a 'first part: differential calculus', a 'second part: integral calculus' and an 'appendix: on differences and series'. The correspondence between these three parts and the three volumes of the large *Traité* is perfect.

There are certain changes in order, for pedagogical reasons: Lacroix did not expect the audience of the *Traité élémentaire* to withstand all the theory of differentiation before seeing any application. There is also a foundational difference: Lacroix wished a 'sufficient degree of rigour and clarity', but without the lengths entailed by certain unnecessary details [Lacroix, 1805, 345–346]. For this reason he decided to use limits instead of Lagrangian power series.

But the main difference, of course, is in depth, as the different intended audiences suggest and the different sizes confirm: 574 8vo pages opposed to 1790 4to of the large *Traité*.

A second edition of the large *Traité* appeared in 1810 (first vol.), 1814 (second vol.), 1819 (third vol.). New developments were included and its bibliography grew considerably (although a new graphical arrangement for the table of contents exaggerated this in terms of number of pages; beware this in Tables 1, 2 and 3 above). Some chapters were subdivided. A few sections were removed because their subjects had become standard in secondary education. But mainly new material was accumulated.

7 IMPACT

It is particularly difficult to assess the impact of Lacroix's *Traité*, as it is not the kind of work a mathematician would normally cite.

However, there is a sort of secondary impact that is noticeable: that of the *Traité élémentaire*. Like the other textbooks by Lacroix, it was hugely successful, having five editions during the author's lifetime, and four posthumous ones, up to 1881. It was translated (at least) into Portuguese (1812, in Brazil), English (1816), German (1817) and Italian (1829). If we take the *Traité élémentaire* as a by-product of the large *Traité*, then the latter must partake of the obvious educational influence of the former.

And there is some evidence also for more direct influence, namely on one of Lacroix's students at the *École Polytechnique*: Cauchy (§25). Grabiner [1981] tries to find the technical origins of Cauchy's rigorous analysis in 18th-century calculus. And for this, she claims, Lagrange's and Lacroix's books were his principal sources [Grabiner, 1981, 79].

One specific example: Cauchy's definition of definite integral comes from 18th-century techniques for approximation of integrals, particularly one by Euler, reported by Lacroix ((7) above). Cauchy not only clearly used it, but he 'consistently used Lacroix's terminology' [Grabiner, 1981, 151]. This path from Euler to Cauchy via Lacroix suggests that Lacroix may have been successful in his main goal: to make the 18th-century calculus, in all its details, much more easily accessible and fruitful to the 19th-century mathematicians.

ACKNOWLEDGEMENT

This article was written while the author was holder of a doctoral scholarship from *Fundação para a Ciência e a Tecnologia*.

BIBLIOGRAPHY

Cajori, F. 1919. *A history of mathematics*, 2nd ed., New York: Macmillan.

Grabiner, J.V. 1981. *The origins of Cauchy's rigorous calculus*, Cambridge, MA: MIT Press.

Grattan-Guinness, I. 1990. *Convolutions in French mathematics, 1800–1840*, 3 vols., Basel: Birkhäuser.

Itard, J. 1991. 'Lacroix, Sylvestre François', in *Biographical dictionary of mathematicians*, New York: Scribner's, vol. 3, 1297–1299.

Lacroix, S.F. 1802. *Traité élémentaire du calcul différentiel et du calcul intégral*, Paris: Duprat. [2nd ed.: 1806, Paris: Courcier. 3rd ed.: 1820, Paris: Courcier. 4th ed.: 1828, Paris: Bachelier. 5th ed.: 1837, Paris: Bachelier. Plus four posthumous editions.]

Lacroix, S.F. 1805. *Essais sur l'enseignement en général, et sur celui des mathématiques en particulier*, Paris: Courcier. [2nd ed.: 1816, Paris: Courcier. 3rd ed.: 1828, Paris: Bachelier. 4th ed.: 1838, Paris: Bachelier. Page references are to the 2nd ed.]

Lagrange, J.L. 1774. 'Sur une nouvelle espèce de calcul relatif à la differentiation et à la intégration des quantités variables', *Nouveaux mémoires de l'Académie* [...] *de Berlin*, (1772), 185–221. [Repr. in *Œuvres*, vol. 9, 439–476.]

Laplace, P.S. 1782. 'Mémoire sur les suites', *Mémoires de l'Académie Royale des Sciences*, (1779), 207–309. [Repr. in *Œuvres complètes*, vol. 10, 1–89.]

Taton, R. 1951. *L'œuvre scientifique de Monge*, Paris: PUF.

Taton, R. 1953a. 'Sylvestre-François Lacroix (1765–1843), mathématicien, professeur et historien des sciences', *Actes du VII^{e} Congrès International d'Histoire des Sciences*, Paris: Acad. Int. d'Hist. Sci. & Hermann.

Taton, R. 1953b. 'Laplace et Sylvestre-François Lacroix', *Revue d'histoire des sciences*, *6*, 350–360.

Taton, R. 1959. 'Condorcet et Sylvestre-François Lacroix', *Revue d'histoire des sciences*, *12*, 127–158, 243–262.

Todhunter, I. 1861. *A history of the progress of the calculus of variations during the nineteenth century*, Cambridge and London: Macmillan.

CHAPTER 21

JEAN-ETIENNE MONTUCLA, *HISTOIRE DES MATHÉMATIQUES*, SECOND EDITION (1799–1802)

Pierre Crépel and Alain Coste

The first wide-ranging history of mathematics for a general readership, this book raised the reputation of the history of mathematics both in its first edition of 1758 and especially in the substantially augmented second edition that is our main concern. Notable is his attention to applied mathematics, even some physics, as well as to pure mathematics.

First publication. 4 volumes, partially ed. J.J. Lalande, Paris: Agasse. Vol. 1, an VII (1798–1799), viii + 739 pages; vol. 2, an VII, 717 pages + errata; vol. III, an X (1802), viii + 832 pages; vol. 4, an X (1802), 688 pages. All vols. with plates.

Photoreprint. Paris: Blanchard, 1968 [with new unpaginated four-page preface by Charles Naux].

First edition. 2 volumes, Paris: Jombert, 1758. xxxvi + 638; 680 pages. Both vols. with plates.

Translations. According to a letter by Montucla, German and Dutch translations were planned [Beaujouan, 1950, 130–131], but they do not seem to have been published.

Related articles: All articles on the 17th and 18th centuries.

1 BIOGRAPHY

Jean-Etienne Montucla was born at Lyon on 5 September 1725 into a family of tradespeople. He studied at the *Collège de la Trinité*, the local Jesuit college, where he took Greek and Latin but also followed an elementary scientific curriculum, mainly under the direction of the astronomer Père Béraud (the college had an observatory), a *correspondant* of the *Académie Royale des Sciences*. Concerning Lyon he recalled especially the memory of old astronomical clock, which one can still see at the cathedral, and of the architectural

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

tour de force of Girard Desargues, namely, a house built on a bridge and overhanging water that was to be destroyed in the 19th century.

Montucla had a flair for languages and privately learned Italian and English, as well as a little German and Dutch, and even some Arabic. His father died in 1741, followed in 1745 by his grandmother who had looked after him. He went to Toulouse to study law, which was not possible at Lyon at the time. He lived later in Paris, where he associated with the intellectuals who met regularly at the house of the bookseller and editor C.A. Jombert, and made his living by working for the *Gazette de France*.

Montucla's first published work was an *Histoire des recherches sur la quadrature du cercle* [Montucla, 1754], part of which is reproduced as an appendix to the *Histoire des mathématiques*. S.F. Lacroix (1765–1843) reissued the book in 1831, augmenting and updating the text. We note that Jombert had already published the work and that the Royal Privilege included this book in the *Histoire des mathématiques*, although this did not appear until four years later. This book on the quadrature of the circle and the support of Jean d'Alembert were responsible for his nomination as a foreign member of the Berlin Academy on 3 July 1755. He later compiled with P.J. Morisot-Deslandes a collection of documents of English origin on inoculation. The publication was in support of the campaign of La Condamine, whose memoir was read at the *Académie des Sciences* on 24 April 1754; however, although approved by the censor on 18 July 1754, Montucla and Morisot-Deslandes's book was not published until 1756, on the occasion of the inoculation of the royal children.

It was in 1758 that Montucla published the two volumes of the first edition of the *Histoire de mathématiques*, and he was already preparing a third on the 18th century. But in 1761 he was appointed to an official position, remote from mathematics, as secretary to the Burser of the Dauphiné, which also obliged him to leave Paris for Grenoble. He benefited from this by getting married in 1763. In 1764–1765 he seems to have spent some time at Cayenne as secretary to the brother of the economist Jacques Turgot. On his return he took up a position at the *Surintendance des Bâtiments* (more or less a ministry of fine arts and architecture) and settled at Versailles, where he lived until the Revolution: he was still in post in July 1792.

During this period, Montucla published with Jombert a new edition of the *Récréations mathématiques* of Jacques Ozanam, which had first appeared in 1694. Montucla did not care much for Ozanam and rewrote a large part of the book. Like the *Histoire des mathématiques*, the *Récréations* deal with many subjects (physics, chemistry and so on). Montucla was not directly credited as the 'author', but he was in later editions and in the English translation of 1803.

In 1784, under the pseudonym 'M.d.C.', Montucla translated and published the third edition of Jonathan Carver's *Travels through the interior parts of North America*, which had appeared in London in 1781, adding notes and a preface. During this period, he continued working on a future reissue of the *Histoire des mathématiques* and on a draft of the Part devoted to the 18th century. Following the Revolution, he had small, badly-paid, official jobs; his request for a pension was granted thanks to his friend the astronomer Joseph-Jérome Lalande (1732–1807), and also Lagrange, but only a few months before his death. He had never been a member of the *Académie Royale des Sciences*, but in the re-

Table 1. Contents of the second edition of Montucla's history. The first column gives the Volume and Part numbers, and the number *of* Books for that Part in the Volume.

Vol., Part; Books	**Page**	**Topics**
1	i–viii	Preface by Montucla.
1, 1; 5	1	Ancient Greeks.
1, 2; 4	350	The Far East.
1, 3; 5	480	The 16th century (with Supplement). [End 739.]
2, 4; 9	1	The 17th century (with Supplement).
2	661	Table of contents; index of vols. 1 and 2.
2	712	Additions, corrections, errata for vols. 1 and 2. [End 718.]
3	v–viii	Preface by Lalande.
3, 5; 4	1	The 18th century: mathematics, optics, mechanics. [End 832.]
4, 5; 5	1	The 18th century: astronomy.
4	584	Supplements: capstan, geography, quadrature, music, antiquity, derivations.
4	661	Notice of Montucla by Leblond, modified by Lalande.
4	673	Table of contents; index of vols. 3 and 4. [End 688.]

organization following year III he was designated on 28 February 1796 as a 'non-resident associate' of the mathematics section of the new *Institut National*.

The new edition of the first two volumes of the *Histoire des mathématiques* appeared in 1799, when Montucla also supervised the printing of a large part of Volume III. He died on 18 December 1799 of a bladder infection, leaving a wife and two children. His friend Lalande, with several assistants, ensured the publication of the third and fourth volumes in 1802. Montucla left among his papers a great mathematical bibliography of more than 600 pages that was never published, although Lalande made use of it in the fourth volume. On his life and work, see especially [Le Blond, 1800; Doublet, 1913] and [Sarton, 1936].

2 CONTENTS OF THE BOOK

It is not easy to find one's way in Montucla's *Histoire des mathématiques*. We consider only the second edition, which encompasses the first. It consists of four 'Volumes' and is divided into five 'Parts' that do not correspond to the Volumes. These Parts are themselves split up into 'Books'. These are also 'appendices', both to the Books and to the Parts. There are summaries (really continuations of the text) of the first four Parts and Books I and II of the fifth, but not of Books III–IX, of the fifth Part. The work also contains two 'prefaces', one 'notice' and a 'table of contents' that today would be called an index. A detailed summary of the whole work may be found on the website http://dalembert.univ-lyon1.fr. Table 1 is a simplified version. We focus mainly on the second edition, but the first edition will also be noted.

3 VOLUME 1 (1758 AND 1799)

A reading of the table of contents shows that the *Histoire des mathématiques* also contains a history of theoretical mechanics, as well as of applied mechanics (clocks, steam engines, and so on), astronomy and optics. In a preface, which he modified in the second edition, Montucla lists his predecessors, with whom he is generally very severe; he particularly accuses them of being biased or giving only a simple chronology. He also describes his intended readership as 'philosophers, professional people, those with a love of the sciences'. This diversity obliges him to supplement his text, both with brief introduction for the less initiated (for example, on algebraic curves) and with appendices to many of the chapters giving details of the proofs for the benefit of professional people.

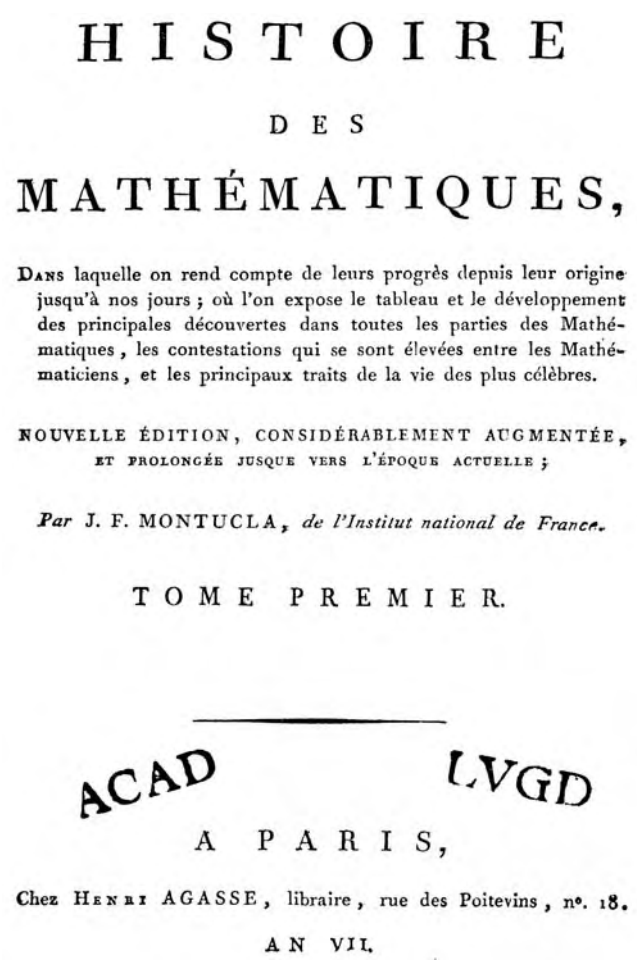
The title page of the first volume of the *Histoire des mathématiques* is shown in Figure 1. The long subtitle promises to give an account of the progress in mathematics from its origin up to the present day, and the first Part of Volume 1 contains a fairly canonical version of the history of mathematics in Ancient Greece. Having read sources available at the time, including Plutarch, Diogenes Läertius and Proclos, he discusses the editions of the great classics: Euclid, Archimedes and Apollonios. While expressing many reservations on the credibility of the little anecdotes that traditionally accompany the history of ancient Greek mathematics, he nevertheless records them faithfully. There is also a lengthy treatment of Greek music appropriate to the fashion in the academics of the 18th century.

The second Part of Volume 1 is at the same time intriguing and deceptive. The intention is to give a history of the mathematics of the Arabs, Persians, Turks, Hebrews, Indians and Chinese. This manifold ambition could not be realized with the sources at his disposal: while his knowledge of Arabic enabled him to correct the transcription of certain headings and proper nouns, the information contained in the Jesuit literature to which he made reference is insufficient, especially in regard to India and China. The whole of this Part is today no more than a curiosity and a testimony to the (lack of) knowledge possessed of the subject matter in the Enlightenment.

The situation is entirely different in the third Part, which deals with the mathematics of the 16th and 17th centuries. Montucla's documentation here is impressive: he seems to have consulted all the books to which he refers, aside from sometimes having seen only a second edition and some problems with the Dutch language. As sources of reference, he relies chiefly on the *Journal des savans*, the *Philosophical transactions* of the Royal Society and the *Mémoires de l'Académie Royale des Sciences*.

The subtitle of the *Histoire des mathématiques* contains a curious phrase: Montucla declares his intention to describe 'the disputes that have arisen between mathematicians' and actually organizes his text with this in mind. Thus, in the last Part of Volume 1, which treats the solution of equations, he devotes several pages to the quarrel between Niccolo Tartaglia and Geronimo Cardano. This Part is dedicated to the history of algebra, and he places particular emphasis on the role of François Viète; while he makes use of John Wallis's *Treatise on algebra* (1685) (compare §2), he is very critical of it.

In the second edition Montucla leaves many chapters unchanged, merely adding various paragraphs, whereby the total number of pages is increased by around 100. For example, the Section on the mathematics of the Orient doubles in length.



HISTOIRE

DES

MATHÉMATIQUES,

DANS laquelle on rend compte de leurs progrès depuis leur origine jusqu'à nos jours ; où l'on expose le tableau et le développement des principales découvertes dans toutes les parties des Mathématiques, les contestations qui se sont élevées entre les Mathématiciens, et les principaux traits de la vie des plus célèbres.

NOUVELLE ÉDITION, CONSIDÉRABLEMENT AUGMENTÉE, ET PROLONGÉE JUSQUE VERS L'ÉPOQUE ACTUELLE ;

Par J. F. MONTUCLA, *de l'Institut national de France.*

TOME PREMIER.

ACAD LVGD

A PARIS,

Chez HENRI AGASSE, libraire, rue des Poitevins, n°. 18.

AN VII.

Figure 1. The title page of the first volume.

4 VOLUME 2 (1758 AND 1799)

The principle of organizing the exposition around 'disputes' between mathematicians is to a large extent maintained: Habakkuk Guldin versus Bonaventura Cavalieri, Gilles Roberval versus René Descartes, and so on. The 'pure mathematical' part of this second Volume is centred around the history of the cycloid, which he humorously calls 'the Helen of the geometers'. He devotes several pages to the competition launched by Blaise Pascal on this subject; and since the edition by Charles Bossut (1730–1814) of Pascal's *Œuvres* (covering the mathematical works) had appeared in 1779, between the two editions of the *Histoire des mathématiques*, Montucla completely rewrote this Section to take account of this (and in particular of Wallis's contribution). The mathematical work of Descartes is described at length and in a fairly positive way, even though Montucla does report the

quarrels with Pierre Fermat and Roberval. Thomas Harriot's contribution to the history of algebraic equations gives him another opportunity to disparage Wallis's treatise.

After this history of algebra in the first half of the 17th century, the first edition continues with a chapter on optics. This is changed in the second edition, which proceeds more logically to the history of mechanics, especially the work of Galileo Galilei on falling bodies and that of Descartes on collisions, of which he is very critical.

As to the history of optics, the links with Joseph Priestley's book *The history and present state of discoveries relating to vision, light and colours* (1772) are close. Priestley wrote with the first edition of Montucla in front of him, and Montucla prepared his second edition after reading Priestley's book, each acknowledging his debt to the other. Montucla devotes two chapters to the 17th-century optics of Johannes Kepler and Descartes, while Priestley focuses on the work of Isaac Newton.

Kepler and Galileo are the heroes of the story of astronomy in the 17th century. The condemnation of the latter enabled Montucla to give vent to an anti-clericism of which traces are to be found throughout the work. He then returns to the pure mathematics of the second half of the century, beginning with a grand eulogy on the scientific work of Wallis, which contrasts with his very negative judgement of the latter's work on the history of mathematics.

But the main concern is the invention of differential calculus in the work of Isaac Newton and G.W. Leibniz. Montucla preserves a prudent neutrality, expressing equal admiration for these two scholars. As to the mechanics of this period, the Huygens–Catalan contention on problems involving centres of gravity is followed by an exposition of the works of the Bernoulli brothers and Leibniz and a very long account of Newton's *Principia*.

As is stated on the title page, the *Histoire des mathématiques* also contains, dispersed throughout the text, a history of mathematicians. For each of these, especially the greatest, who worked in many areas, the question arises of where to put the biography. Montucla's solution seems to have been to seek a balance whereby the biographical notes neither cause too much disruption of the chronological sequence of the history of ideas, nor follow too closely on one another. Thus, one finds the notes on Descartes, Newton and Leibniz in the pure mathematical Sections, those on Kepler and Galileo in the chapters on astronomy, and that on Christiaan Huygens in the chapter on mechanics. These notes are rather brief and not of great use today.

5 VOLUME III (1802)

The third and fourth Volumes, together labelled 'Fifth Part', comprise a history of 18th-century mathematics; it caused Montucla new difficulties. In the 18th century, mathematicians grew in number and professionalism, and the memoirs published increased in line with the number of journals containing them. To the old ones (*Journal des savans*, *Mémoires de l'Académie Royale des Sciences* and *Philosophical transactions*) were added the *Mémoires* of Berlin, Turin, Saint Petersburg, Göttingen, and so on. Deciding what is important among the works of one's own period is always a difficult task. Someone virtually self-educated like Montucla comes up against problems other than those arising in the work of earlier centuries, even though he consulted the professionals.

Volume 3 is divided into four Books, of which the first contains the history of mathematics in today's sense and makes up more than half of the Volume. It begins by describing the state of research on the solution of algebraic equations and the obstacles apparently encountered by the methods of A.T. Vandermonde and J.L. Lagrange and the calculus of symmetric functions beginning with the equation of the fifth degree, and the resulting interest in the approximate solution of Newton, as well as those of Johann Bernoulli, Brook Taylor and Lagrange. He then connects up with one of his favorite themes, the theory of curves and applications of the differential calculus, a subject that enables him to study the Newton-Leibniz quarrel and also the attacks of Michel Rolle and George Berkeley and the defences of Joseph Saurin, Benjamin Robins and Colin Maclaurin.

As he approaches his own era, Montucla changes his method of exposition. Conscious of not having mastered all the references, he divides mathematics up into a number of different areas (he actually notes the beginnings of specialization inside mathematics, even though the same mathematicians crop up in different areas). In each of these areas, he chooses one or two general (rather than original) treatises to form the basis of his text, augmenting these from time to time with references to various memoirs. In the case of differential equations, for example, he makes explicit use of the *Traité de calcul intégral* of Louis Bougainville (1754–1756) and the *Eléments de calcul intégral* of Thomas Leseur and François Jacquier (1768). The former enables him to recall the works of D'Alembert and the latter those of Leonhard Euler, but he admits his difficulties in understanding and summarizing the recent work of the Marquis de Condorcet! He then turns to the theory of series. After describing in detail the classical works of Newton, Leibniz and the Bernoullis, even giving some explicit calculations, he mentions the recurrent series of Abraham de Moivre and continued fractions. As to more recent work, Euler serves as his guide.

Montucla tries to give a survey of other areas of research, but it is clear that he has no longer been able to read everything, and only cites the Italian mathematicians at second hand. For the calculus of finite differences, he quickly returns to Bossut's and J.J. Cousin's recent treatises and the articles in the *Encyclopédie méthodique*.

Lagrange's *Théorie des fonctions analytiques* (1797) is Montucla's source for the theory of functions. He departs at this point from the exposition of new mathematical theories to spell out several geometric applications that arise, for example, in the exchanges between Bernoulli and Leibniz on geodesic problems and orthogonal trajectories. Montucla thus returns with a sense of relief to the universe of controversies between mathematicians that had formed the stylistic setting of the previous Volume. The discussions between Johann and Jacob Bernoulli occupy several pages.

It is at this point that matters become complicated owing to Montucla's death. The first chapter that follows is the one on partial differential equations; Lalande did not feel himself capable of reviewing this difficult topic and turned to Lacroix for help. The chapter is centred around the discussion of priorities between Euler and D'Alembert, and it is not known whether it is Montucla or Lacroix himself who refers to the *Traité de calcul différentiel et intégral* of the latter (1797–1800) as being 'the newest and most complete' authority on this question (§20). The Section on the calculus of variations contents itself with a reference to Lagrange. The last two topics are logarithms (inspection of the tables and the debate on logarithms of negative quantities) and probability (D'Alembert's reser-

vations on the foundations of the theory and their defence by Daniel Bernoulli), along with its numerous applications to questions of finance, census, medicine, and so on.

It was Fortia d'Urban who supervised the editing of the pages of 'Book II' on optics: in his notes, he gives some idea of the state of the manuscript at the time of Montucla's death and of what he has done to prepare it for publication. The table of contents at the beginning is Montucla's, but many of the proposed articles were only partially written up. Thus the very long Section V on achromatic lenses amounted to only 10 pages; Fortia adds another 40, mainly on the work of D'Alembert. To round off Section IV, Fortia transcribes the relevant passages from Volume II of the *Récréations mathématiques* retaining Montucla's style. Beginning with Section VI of Book II, the editing is again taken over by Lalande, apparently with fewer scruples in regard to Montucla: he changes the proposed titles in cases where they seem somewhat inconvenient to him and makes frequent reference to his own works ('as I have in my *Ephémérides*'). There is also a Section on the construction of telescopes, spectacles, microscopes and other instruments without much in the way of theory. The last Sections bring together a large number of notes on all kinds of inventions, phantasmagoria, optical spectacles, the oculer harpsichord of Père Castel, and so on; all this work might be found in a 19th-century magazine on the popularization of science.

It seems likely that 'Book III', on mechanics and fluid mechanics, was written by Montucla himself. The author recognizes that he has a perfect guide to the area in Lagrange's book *Méchanique analitique*, which appeared in 1788 (§16): 'there could be no surer guide or more profound historian, leaving nothing to be desired in the way of history or scholarship'.

Montucla begins with the principles of statics followed by those dynamics, giving pride of place to explaining the principle of conservation of live forces (*forces vives*, attributed chiefly to Huygens) and D'Alembert's principle. The debate on live forces, which stimulated scientific fashions for 40 years, is a choice morsel for Montucla: its history take up 50 pages, of which the more theoretical part concludes with the controversy between P.L. Maupertuis and J.S. Koenig on the principle of least action. As to applications, he chooses the problem of tautochrones in resistant media, vibrating strings and the discussions between D'Alembert, Euler, Daniel Bernoulli and Lagrange, and finally ballistics via the work of Benjamin Robins annotated by Euler. Fluid mechanics is treated more briefly, the author referring to the *Hydraulica* of Johann Bernoulli, the *Hydrodynamica* of Daniel Bernoulli and the *Traité des fluides* of D'Alembert. He cites, without giving technical details, the new methods (partial differential equations) of Alexis Clairaut, D'Alembert and Lagrange (without mentioning Euler!). A long Section on the hydraulics of rivers borrows directly from Bossut many pages on the Italian hydraulic engineers, augmented by several allusions to the works of Forest de Bélidor.

The fourth Book of Volume III deals with machines and is written by Lalande. After mentioning human forces and those of horses, and frictional resistances as studied by Guillaume Amontons, G. Riche de Prony and C.A. Coulomb, he goes on to describe actual machines: Marly machines and various pumps, steam engines, windmills, steamships and the first trials of Joseph Montgolfier. He ends with clocks and automata. This Book is rather untidy throughout.

6 VOLUME 4 (1802)

Volume 4 contains hardly any pure mathematics but much astronomy, which explains Lalande's haste in completing Volume III: to have a greater input in the text. The references to his own *Astronomie* (editions from 1764) are numerous, and one also finds signed notes and a record of events that took place after Montucla's death.

The fifth Book is the one concerned with astronomy as such and covers the expected topics: the system of the world, the stars, and the theory of the Sun. Much space is devoted to the theory of the Moon and the relevant works of Clairaut, D'Alembert, Euler, Edmund Halley and John Machin, among others. Eclipses are also studied, along with the solar transits of Mercury and Venus. The highlight of this Book is the discovery of a new planet (Uranus) by William Herschel, who is the leading light in Book V.

Book VI deals with physical astronomy. It moves away from mathematics to some extent, although there are links with refraction and the force of the Earth. But the long history of the measurement of the arc of the meridian at the pole and at the equator to verify the flattening of the Earth sometimes contains the story of a journey or even a new item (the murder of Semiergue). It then returns to the theory, as developed by Clairaut, MacLaurin, D'Alembert and R.J. Boscovich, before analyzing at length the *Mécanique céleste* of Laplace. This difficult work only began to appear in 1799 (§18), so there is no doubt that the account is due to Lalande. The Book then returns to classical topics: the aberration highlighted by the experiments of James Bradley, the precession of the equinoxes and the nutation of the axis of the Earth following the works of D'Alembert. For the obliquity of the ecliptic, the author again follows Laplace. As to the means of Jupiter, Saturn and Uranus, frequent reference is made to the tables of Lalande, and the same goes for the pages relating to the story of the return of Halley's comet in 1759.

Book VII is a collection of astronomical tables and calendars, and the last two Books (VIII and IX) bring together a large number of questions more or less related to navigation. In Book VIII, these are 1) the construction of boats, their stability, the oars, the sails, and so on, in which the author analyses the French students of these questions (for example, Pierre Bouguer, Sébastien Vial du Clairbois and Charles Romme) as well as the French adaptation of the book by Frederick Chapman; and 2) the maneuvering of boats, the theory of steering, swaying and pitching, and the way boats are loaded, which are rather more mathematical: Euler, Clairaut, Bouguer, J.C. Borda and Bossut contributed to the theory. The analysis of Don Jorge Juan's treatise, the *Examen maritime* (1771), sums up recent progress on these questions.

Book IX, the last, deals with the location of ships: compasses and the calculation of speeds. At the very end there is a summary of the various methods for solving the problem of longitude, finishing up with the story of John Harrison, which tones down the rather too exclusively French aspect of this recent history of navigation.

The Volume ends with six appendices. The first reproduces a report by Borda on a capstan project of Charles de la Lande. The second is a very brief history of geography, or rather a history of the discoveries of the navigators with remarks on the Atlantic and on the discovery of America by the Vikings before Columbus. There is also some information about maps, in particular a page on the ancient chart of Peutinger. The third appendix, written by Montucla, is a history, based on the book published in 1754, of attempts to

square the circle. The fourth appendix returns to music in antiquity and the fifth includes several pages on ancient philosophers. The sixth gives a summary of L.F.A. Arbogast's book *Du calcul des dérivations* published in 1800. The work ends with a biography of Montucla arranged by Lalande in 1800 from that of Guillaume Le Blond.

7 RECEPTION OF THE WORK

Montucla's *Histoire des mathématiques* was circulated widely and both editions quickly gained a great reputation. It has been paraphrased and even plagiarized, and at least used as a source on innumerable occasions in the course of the last 250 years. However, we know of no systematic study of its reception and utilization.

7.1 Reception and utilization of the first edition. It has the distinction of being cited before its publication in the most celebrated work of the 18th century, the *Encyclopédie*; it was first mentioned by D'Alembert in his article on 'Géométrie' in volume 7 (1757). It is used later in over a dozen articles of various types, such as 'Logarithme' and 'Optique', both by D'Alembert and the chevalier de Jaucourt. Other passages are copied verbatim in Yverdon's *Encyclopédie* (a Swiss modification of the *Encyclopédie*) and in the *Supplément*, such as the article 'Conique'.

Reaction in the journals was immediate. Beginning in 1759, long accounts, rather banal yet very favourable, appeared in the *Journal des Savans* (259–267, 467–474), the *Journal de Trévoux* ((1759), 489–512, 1759–1791, 2501–2527 and (1760), 122–150), and the *Mercure de France* (January 124–138, February 111–117). Thus the book was ignored by nobody, from Bossut through Condorcet to Lagrange, and everyone made use of it.

7.2 Immediate reception of the second edition. The second edition likewise did not pass unnoticed. The *Journal de Paris*, a daily newspaper, reported the issue of Volumes 1 and 2 in its edition of 15 *thermidor* of year VII and of Volumes 3 and 4 that of 25 *prairial* of year X, but only in simple announcements of a few dozen lines. It even gave the price: 31 francs 50 for the first two and 31 francs 30 for the other two. A significant coincidence is worth mentioning: 1802 also saw the publication of the *Essai sur l'histoire générale des mathématiques* by Bossut; this work received much longer reviews in the *Journal de Paris*, and also in the *Décade philosophique*. And, of course, the journalists had to compare the two histories, to Bossut's advantage in both cases. They recalled the strengths and weakness of the first edition of Montucla and the criticism it had suffered in the intervening 40 years, and pointed out the differences in methodology of the two authors. They were assisted in this by Bossut himself, who devoted three pages to the subject in his own preface (pp. v–vii), adding that he had not yet seen the second edition. As an example here is a passage from the *Journal de Paris* for the 3rd complementary day of year X (1800–1801):

> There is no detailed history of mathematics like that of Montucla. Citizen Bossut informs us that his object is different. In each branch of the sciences, he considers only the fundamental ideas and their major consequences. As a result, the picture he paints is infinitely more effective, since in a convincing way it brings together into one panorama the most magnificent of all the sights,

> the origin of this kind of knowledge, that certainty always accompanies, and the successive progression of discoveries that have almost completely changed the face of the earth.

One is led into the belief that the mathematician Bossut had a greater mastery of the subject than a well-informed amateur like Montucla.

7.3 Subsequent evaluation and utilization. The history of subsequent evaluations comes more into line with what we have said in our analysis. Whatever they thought of it, historians of science from Moritz Cantor to George Sarton took it very seriously. In particular, Volumes III and IV, and even segments of Volume II, enjoy a rather special status: they act as a kind of intermediary between primary and secondary sources.

The *Biographie universelle* of Abbot Feller (1849 edition, volume 6) shows no hesitation in confirming (doubtless for ideological rather than historical reasons) that 'the last two volumes, printed after the death of the author under the direction of Lalande, more often than not provide only a heavy overview of optics and physical astronomy, where one sometimes finds random judgements'. Finally, in our times, the historian Kurt Vogel has rightly noted that 'Montucla had no successor until Moritz Cantor' [1971, 501]. The fact remains that Montucla's *Histoire des mathématiques* continues to be read and used up to the present day. After all, as we saw in the publication history, there is a modern photoreprint.

BIBLIOGRAPHY

Beaujouan, G. 1950. 'Lagrange et Montucla', *Revue d'histoire des sciences, 3*, 128–132.

Bossut, C. 1802. *Essai sur l'histoire générale des mathématiques*, 2 vols. Paris: Louis.

Cantor, M. 1880–1908. *Vorlesungen über Geschichte der Mathematik*, 4 vols., Leipzig: Teubner.

Doublet, E. 1913. 'Montucla—l'historien des mathématiques', *Bulletin de l'Observatoire de Lyon, 5* (December), 2–8.

Evieux, A. 1926. 'Un mathématicien lyonnais : Jean-Etienne Montucla (1725–1799)', *Le Salut Public*, vendredi 16 avril, 5. [Daily Lyon newspaper, available in the *Bibliothèque municipale de Lyon*, callmark 950 001.]

Le Blond, A.S. 1800. *Notice historique sur la vie et les ouvrages de Montucla*, Versailles: Société Libre d'Agriculture de Seine-et-Oise.

Montucla, J.E. 1754. *Histoire des recherches sur la quadrature du cercle*, Paris: Jombert.

Sarton, G. 1936. 'Montucla (1725–1799). His life and works', *Osiris, 1*, 519–567 [with portraits and autographs].

Swerdlow, N. 1993. 'Montucla's legacy: the history of the exact sciences', *Journal of the history of ideas, 54*, 299–328.

Vogel, K. 1971. 'Montucla, Jean-Etienne', in *Dictionary of scientific biography*, vol. 9, 500–501.

CHAPTER 22

CARL FRIEDRICH GAUSS, *DISQUISITIONES ARITHMETICAE* (1801)

O. Neumann

The *Disquisitiones arithmeticae* defined in an authoritative way the substance and methods of number theory (and also, in part, of the theory of equations) for the subsequent five or six decades of the 19th century. It contained the first proof of the reciprocity law for quadratic residues, an entirely new approach to the theory of binary quadratic forms and, for the first time, a general, coherent and explicit theory of the equation $x^n - 1 = 0$ in the theory of cyclotomy (to use a later name).

First publication. Leipzig: Fleischer, 1801. xvii + 668 + [12] pages.

Further editions. As *Werke*, vol. 1 (ed. E. Schering), Leipzig and Berlin: Teubner, 1863. 2nd revised ed. 1870 (photorepr. Hildesheim: Olms, 1973), 1–463.

French translation. Recherches arithmétiques (trans. A.-Ch.M. Poullet-Delisle), Paris: Courcier, 1807.

German translation. *Arithmetische Untersuchungen* (trans. H. Maser), in Gauss, *Untersuchungen über höhere Arithmetik*, Berlin: J. Springer, 1889 (photorepr. New York: Chelsea, 1965), 1–453.

Russian translation. In *Trudi po teorii chisel*, Moscow: Russian Academy of Sciences, 1959.

English translation. *Disquisitiones arithmeticae* (trans. Arthur A. Clarke), New Haven: Yale University Press, 1966. [Rev. ed. 1986.]

Spanish translation. Disquisitiones arithmeticae (transl. H. Barrantes Campos, M. Josephy and A. Ruiz Zúñiga), Santa Fe Bogota, D.C.: 1995.

Catalan translation. *Disquisicions arithmètiques* (trans. G. Pascual Xufré), Barcelona: 1996.

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

Related articles: Dirichlet (§37), Weber (§53), Hilbert on number theory (§54), Hilbert on mathematical problems (§57), Dickson (§65).

1 INTRODUCTION

September 1801 saw the publication in Leipzig of a scientific work, written in Latin, with the modest title *Disquisitiones arithmeticae* (hereafter 'D.A.'; in English the title is 'Arithmetical enquiries'). It was quickly recognized by contemporary experts, especially in France, as a masterpiece of unprecedented organization, rigour and extent, which transformed number theory from a scattering of islands into an established continent in mathematics.

Its author was the 24-year-old Carl Friedrich Gauss (1777–1855). After studying mathematics in Göttingen from 1795 to 1798 Gauss lived as a private scholar in Braunschweig. He had conceived the idea of writing such a book while still a student in 1796. The execution of this plan was to be deferred until 1800. The most important source for the history of the writing of D.A. is the mathematical diary kept by Gauss from 1796 to 1814, which contains a total of 146 entries [Gauss Diary]. The original manuscript of the book is not extant. A first draft, with the title 'Analysis Residuorum' and having the theory of congruences as a key feature, was completed from two fragments discovered by Uta Merzbach in 1975 and published in Volume 2 of the *Werke* [Merzbach, 1981]. However, Sections 4 and 5, which form more than half of the printed version, are missing from this first draft.

We shall take for granted a familiarity with the most significant dates in Gauss's life and work. The reader may compare the lists in [Küssner, 1979, 11–12] and [May, 1972], as well as the letters compiled in [Biermann, 1990]; see also §23.1.

2 WHAT IS THE *DISQUISITIONES ARITHMETICAE* ABOUT?

Here is an extract from the Foreword (with elaborations indicated in square brackets):

> The investigations described in this book have to do with that part of mathematics which deals with whole numbers [namely, $0, 1, -1, 2, -2, 3, -3, \ldots$], with fractional numbers being largely ignored and imaginary numbers [that is, the irrational complex numbers] left out altogether [...] thus the whole numbers (and fractions, insofar as they are defined in terms of whole numbers) form the subject of arithmetic [...] so it would seem appropriate to distinguish two branches of arithmetic, with the above ['art of counting and calculating'] regarded as belonging to elementary arithmetic, while all general considerations of the specific relations of whole numbers are a part of higher arithmetic ['arithmetica sublimior'] that will be our sole concern here.

Gauss accordingly divided his D.A. into seven Sections (see Table 1). Below 'art. 25', say, will refer to Article 25.

Gauss also noted in the Foreword (and as a reference in art. 44) an eighth Section 'which has already been mentioned in a few places in this volume and contains a general treatment of algebraic congruences of every degree'. Among his papers there is indeed a fragment

Table 1. Contents by Sections of Gauss's book.

Sec.; Page	Arts.	Short 'Title' or Description: other included topics
Dedication to Duke of Brunswick–Luneburg (3 pages). Preface (6 pages).		
I; 1	1–12	'Congruences of integers in general': least residues.
II; 8	13–44	'Congruences of first degree': uniqueness of prime number decomposition, polynomials with integer or rational coefficients.
III; 41	45–93	'Power residues': primitive roots, Fermat–Euler theorem.
IV; 92	94–152	'Congruences of the second degree': quadratic reciprocity law.
V; 165	153–307	'Forms and indeterminate equations of the second degree': binary and ternary quadratic forms with integer coefficients, proper and improper equivalence, composition of binary forms and classes, genera.
VI; 540	308–334	'Various applications of the above': partial fractions, decimal fractions, primality tests, factorization.
VII; 592	335–366	Cyclotomy: periods, solution by radicals, regular polygons. [End 665.]
Additions (3 pages), Tables (6 pages), Errata (4 pages).		

with the title 'Caput octavum' ('eighth Section': *Works*, vol. 2, 212–242), which was completed for publication.

Among Gauss's predecessors, L. Euler (1707–1783), J.-L. Lagrange (1736–1813) and A.-M. Legendre (1752–1833) had already dealt to some extent with certain types of Diophantine equations and achieved considerable success [Weil, 1984, chs. 3–4]. Sections 3–6 of D.A. make brief mention of the work of these authors, especially the comprehensive *Essai* [Legendre, 1798], but they contain far more general results using new methods.

Sections 1–4 of D.A. deal exclusively with whole numbers. Their content would now no longer be described by the Gaussian term 'higher arithmetic', but rather as 'elementary multiplicative number theory'. An important exception is the so-called 'Gauss lemma' on polynomials, which asserts (in a modern formulation) that the product of two primitive polynomials (in one unknown with integer coefficients) is again primitive (art. 42).

The fifth Section specifies the Gaussian definition of 'higher arithmetic' more precisely, comprising a study of 'binary and ternary quadratic forms', and thus of homogeneous polynomials of degree two in two or three unknowns, with integer coefficients. Here Gauss exhibits in detail a very significant and methodical self-restraint. Every binary quadratic form can be decomposed into linear factors:

$$\begin{aligned} ax^2 + 2bxy + cy^2 &= a^{-1} \cdot \left[(ax+by)^2 - (b^2-ac)y^2\right] \\ &= a^{-1} \cdot \left(ax + by + \sqrt{(b^2-ac)}y\right) \cdot \left(ax + by - \sqrt{(b^2-ac)}y\right). \quad (1) \end{aligned}$$

If, for a, b, c as integers, $(b^2 - ac)$ is not a square, then $\sqrt{(b^2-ac)}$ is irrational. Euler, Lagrange and Legendre had used these linear factors to good effect without developing a fully-fledged theory of quadratic forms. Gauss first became acquainted with the work of

these authors in Goettingen in October 1795. In Section 5 of D.A. he decided to dispense (almost) completely with the irrational linear factors and to work instead with identities between polynomials with integer coefficients. This naturally makes an understanding of Section 5 much more difficult for later readers. It also contains his treatment of continued fractions.

Why did Gauss adopt this course? Felix Klein (1849–1925) suggested that he had applied the linear factors and their representation by a lattice in the plane to a heuristic purpose [Klein, 1926, 38–40]. There is certainly no evidence for such a conjecture in Gauss's papers and letters, as Klein admitted. It seems likely that in the long run Gauss was moved to discard quadratic irrationalities by noticing the deficiencies and difficulties in Lagrange and Legendre.

Section 6 consists partly of applications of Sections 1–4, for example to partial fractions and periodic decimals, and partly of applications of Section 5 to the factorization of integers.

Section 7 deals with the equation $x^n = 1$ and its solutions in the domain of complex numbers. These solutions are regarded without further comment as points of the complex plane. They all lie on the unit circle and divide its circumference into n arcs of equal length, whence the name (due later to J.J. Sylvester) 'theory of cyclotomy'. Gauss explicitly informs the reader here that, in as much he is applying (algebraic) complex numbers, he is stepping outside the domain of higher arithmetic, but establishing a narrow connection with this domain.

In the unpublished eighth Section, Gauss again goes beyond the subject of higher arithmetic as originally defined. He considers polynomials in one variable with integer coefficients and congruences between such polynomials modulo a fixed prime p or modulo a pair $(p, f(x) \mod p)$, where $f(x)$ is irreducible mod p. In modern terminology, with $\boldsymbol{F}_q$ denoting the field of $q = p^n$ elements, Gauss proved that the polynomial ring $\boldsymbol{F}_p[x] = \boldsymbol{Z}[x]/(p)$ is Euclidean, and he investigated the factor rings $\boldsymbol{F}_p[x]/(f(x) \mod p) = \boldsymbol{Z}[x]/(p, f(x))$. When $f(x) \mod p$ is irreducible, this factor ring is a finite field $\boldsymbol{F}_q$, and $f(x) \mod p$ is a divisor of $(x^{q-1} - 1) \mod p$. This led to the theory of finite fields of characteristic p, which might be called 'the theory of cyclotomy modulo p'. This theory was then developed independently of Gauss by Evariste Galois (1811–1832) in a paper published in 1830, although it made appeal to Gauss's concept of congruence in its theory of 'imaginary roots' of congruences.

Gauss himself later introduced algebraic complex numbers as a subject within higher arithmetic, in which he investigated the numbers $a + bi$ (a and b as whole numbers) and thus proposed an 'extension of the field of arithmetic' [Gauss, 1825, 1831]. The subsequent development of number theory followed up this extension with great vigour.

After Gauss, L. Kronecker (1823–1891) used D.A. as a model for the most consistent subject definition of arithmetic. He appealed to Gauss in formulating the programme of 'general arithmetic', which discarded algebraic irrationals, described a theory of polynomials in arbitrarily many unknowns with integer coefficients (and their congruences), and included the theory of algebraic functions. One can indeed replace, for example, the irrational number $\sqrt[3]{2}$ by the residue class of $X \mod (X^3 - 2)$.

3 CONTENT AND RECEPTION OF SECTIONS 1–4

In Section 1 of D.A. Gauss introduced the relation of congruence,

$$a \equiv b \pmod{m} \quad \Leftrightarrow \quad m \text{ divides } (a-b), \tag{2}$$

for integers m, a and b. The suggestive symbol '$\equiv$' should emphasize the similarity with the relation of equality. The Latin word 'modulus' means 'measure' or 'scale'. The term 'module (over a ring)', familiar in modern mathematics, arose from it via a sequence of changes of meaning.

The idea behind congruences had been known and exploited for a long time: residue classes mod m were familiar as 'arithmetic progressions with common difference m', and congruences were used as a tool in the solution of equations in integers. The notion of congruence nevertheless exerted a strong influence over a long period. It was extended step by step to algebraic numbers and polynomials by Gauss (in Section 8 of D.A.) and many others, including C.G.J. Jacobi (1804–1851), E. Kummer (1810–1893), J.P.G. Lejeune-Dirichlet (1805–1859), G. Eisenstein (1823–1852) and Kronecker. R. Dedekind (1831–1916) led the way in shifting the focus from the relation $a \equiv b \pmod{m}$ to the set

$$M = \{x \mid x \equiv 0 \pmod{m}\}, \tag{3}$$

which he likewise called a 'module'; knowledge of it is logically equivalent to a knowledge of the congruence relation:

$$a \equiv b \pmod{m} \quad \Leftrightarrow \quad (a-b) \in M. \tag{4}$$

Dedekind's generalization was in the use of the term 'module' to designate any non-empty set M of complex numbers with the property that 'from $x, y \in M$ it follows that $x - y \in M$', and hence to any additive subgroup of $\boldsymbol{C}$. He developed a self-contained calculus for these modules with four operations (sum, product, intersection and quotient) and thus obtained on the one hand a powerful tool in the theory of divisibility of algebraic numbers, and on the other the even more general concept of 'dual group', or 'lattice' in modern terminology. It should be mentioned that the notion of module lent conceptual clarity to Gauss's composition of quadratic forms in Section 5 of D.A. In all aspects of his theory of algebraic numbers, Dedekind dealt first with modules and then with the special case of ideals.

3.1 The following 'theorem on the ' was proved in Section 3 (art. 55):

If p is a prime, then there is a number g such that

$$1, 2, \ldots, (p-1) \mod p \text{ coincide with } 1, g, g^2, g^3, \ldots, g^{p-2} \mod p. \tag{5}$$

Any such g is called 'a primitive root modulo p'.

3.2 The following conjecture of Emil Artin (1898–1962) is not yet completely settled: for any number $a \neq -1$ that is not a square there are infinitely many primes p such that a is a

primitive root modulo p. The number of such primes less than x is given by an asymptotic formula in x.

This conjecture, dating from 1927, was made more precise about 35 years later by H. Heilbronn (1908–1975) using extensive numerical calculations to obtain an improved asymptotic formula, which was finally proved in 1967 by C. Hooley under the assumption of the still unproven generalized Riemann hypothesis.

3.3 The theory of power residues begins in Section 3 of D.A. (art. 60). Let p be a prime, $n \geqslant 2$, m the greatest common divisor of n and $p-1$, and the number a not divisible by p. Then we have the following consequence of the theorem in Section 3.1 above:

a is an nth-power residue modulo $p \Leftrightarrow \sqrt[n]{a} \mod p$ exists $\Leftrightarrow x^n \equiv a \pmod p$ is soluble $\Leftrightarrow$ a is an mth-power residue modulo $p \Leftrightarrow a^{(p-1)/m} \equiv 1 \pmod p$ (Euler's criterion).

Therefore, without loss of generality one can restrict attention to the case $n|(p-1)$ or $p \equiv 1 \pmod n$, since only then is $m = n$. The central and difficult question on nth-power residues is as follows (in a form barely within elementary number theory): for which of the infinitely many primes p with $p \equiv 1 \pmod n$ is the congruence $x^n \equiv a \pmod p$ soluble for a given a?

3.4 The special case $n = 2$ of quadratic residues includes all primes (since $p \equiv 1 \pmod 2$ for $p \neq 2$) and had already been thoroughly investigated by Euler, Lagrange and Legendre, as Gauss learnt in October 1795 on his matriculation in Göttingen. He studied them exhaustively in Section 4 of D.A. The fundamental theorem ('theorema fundamentale'), or reciprocity law (following Legendre, 'loi de réciprocité') for quadratic residues, is stated in art. 131: 'If p is a prime of the form $4n+1$, then so will be $+p$, however, if p is of the form $4n+3$, then $-p$ will be a residue or non-residue for every prime which is a residue or non-residue of p'. Note that here Gauss is choosing from $+p$ and $-p$ the number

$$p* := (-1)^{(p-1)/2} \cdot p \equiv 1 \pmod 4, \tag{6}$$

that is, a number in the sequence $-3, +5, -7, -11, +13, +17, -19, \ldots$. The law can be expressed symmetrically as follows: $p*$ is a residue of $q \Leftrightarrow q$ is a residue of p; p^* is a non-residue of $q \Leftrightarrow q$ is a non-residue of p, where q is a positive odd prime $\neq p$.

The question mentioned above has the following answer in the quadratic case. The congruence $x^2 \equiv a \pmod p$ is soluble if and only if p lies in certain residue classes mod $4a$. These classes form a multiplicative group.

The choice of p^* is an important canonical standardization which crops up again in the theory of higher-power residues; it was later used by Gauss himself in the study of biquadratic (that is, fourth-power) residues in the paper [Gauss, 1825, 1831]. In the light of these later developments, p^* could be called a 'primary number' (or, in the context of quadratic number fields, a prime discriminant). p^* also occurs in the theory of cyclotomy, namely, as the square of the difference of the two so-called periods of $(p-1)/2$ terms (arts. 356–357). Thus, every square root is an integral linear combination of roots of unity. This leads to new but non-elementary proofs of the reciprocity law, as Gauss already knew.

3.5 Who or what in Braunschweig had prompted Gauss to occupy himself with the subject of Sections 1–4 of D.A.? Very little is known about this. Küssner [1979, 35ff.] has drawn attention to the fact that the library of the *Collegium Carolinum*, which Gauss attended from 1792 to 1795, contained a copy of the *Treatise of algebra* (London 1685) by John Wallis (1616–1703). However, it did not have the works of P. de Fermat (1601 or 1607/08?–1665), Euler, Lagrange or Legendre. But the *Treatise* could have confirmed Gauss to study number theory. It deals inter alia with the rational solutions of the equation $t^2 - du^2 = 1$, where d is a positive non-square number, and an unsuccessful attempt to prove that this equation always has a positive integer solution. Wallis is mentioned in art. 202 of the D.A. Wallis had already tried to interpret geometrically the square roots of negative numbers.

Gauss became interested very early on in the factorization of large numbers. It is known from his papers that he also carried out such factorizations using the quadratic forms $ax^2 + cy^2$ [Maennchen, 1918, esp. arts. 6–7]. But unfortunately we do not know how he came to use quadratic forms for this purpose. I believe that Gauss had come across the very old formula

$$\left(a^2 + b^2\right)\left(c^2 + d^2\right) = (ac - bd)^2 + (bc + ad)^2 \tag{7}$$

in the early days of his youth in Braunschweig (in the literature? in Wallis?), especially as applied to the unit circle, that is, with

$$a^2 + b^2 = c^2 + d^2 = 1, \tag{8}$$

which embraces the addition formulae for sines and cosines. It might further be suggested that one day he read this formula from right to left and asked himself the question: when is a factor m of a number $x^2 + y^2$ again the sum of two squares, that is, $m = s^2 + t^2$? It is immediately clear that only the case when the greatest common divisor $\gcd(x, y) = 1$ is of interest, and that in this case examples show that the answer is in the affirmative. One thus obtains an efficient method for factorizing the sum $x^2 + y^2$. It is a relatively short step from this sum to sums of the form $ax^2 + cy^2$, for which one also has a product formula

$$\left(ax^2 + cy^2\right)\left(aX^2 + cY^2\right) = (axX - cyY)^2 + ac(xY + Xy)^2, \tag{9}$$

which appears in more general form in arts. 154 and 229 of D.A. It is also conceivable that the young Gauss classified numbers n according to the difference $n - ([\sqrt{n}])^2$, for it is well known that one only needs to check numbers $\leqslant \sqrt{n}$ as possible factors of n.

Gauss conjectured the reciprocity law after making extensive numerical observations on the factorization of $A^2 + k$ ($k = 1, 2, 3, 4$ and further values) in the period up to the beginning of 1795 (see the foreword of D.A. and [Maennchen, 1918]). This is evident from a memorandum made by Gauss in this period and published in [Biermann, 1977], on the odd prime factors of $A^2 + 1$ and $A^2 + 4$, which, as he suggests, coincide precisely with the primes of the form $4n + 1$.

The reciprocity law had already been stated by Euler and Legendre, and was formulated independently by Gauss in March 1795. He found his first proof in April 1796; it forms the centrepiece of Section 4 of D.A., thus consummating elementary number theory in a definitive fashion. We find a second proof in Section 5 that goes beyond elementary

number theory and rests on the theory of quadratic forms. Gauss subsequently gave six further proofs: the first has been applied by J. Tate in algebraic K-theory [Milnor, 1971, art. 11].

Ferdinand Minding (1806–1885) compiled the first German textbook, *Anfangsgründe der höheren Arithmetik* (1832) in which Sections 1–4 of D.A. were popularized. The first Russian textbook of number theory, 'The algebra and calculus of finite quantities' by N.I. Lobachevsky (1792–1856), appeared in Kazan in 1834; it was based around Gauss and Legendre. The doctoral thesis (roughly corresponding to the German *Habilitationsschrift*) 'Theory of congruences' by P.L. Chebyshev (1821–1894), which appeared (in Russian) in Saint Petersburg in 1849, followed Sections 1–4 of D.A. and gave a critical appraisal. Chebyshev, and H.J.S. Smith (1826–1883) in [Smith, 1859–1865, art. 16], were apparently the first to accredit Euler with priority in stating the reciprocity law.

The reciprocity law for quadratic residues provided a challenging model for the investigation of higher-power residues, that is, nth-power residues for $n \geqslant 3$. These investigations extended beyond elementary number theory, since they required the use of nth roots of unity and thus the contents of Section 7 of D.A., as Gauss had already indicated in his work [1825, 1831] on biquadratic residues. Up to around 1860, the generalization of Gauss's results had been regarded as the main aim of number theory. He himself settled the case $n = 4$ in pioneering fashion by proving the uniqueness of prime factorization for the ring $\boldsymbol{Z}[i]$ consisting of the complex numbers $a + bi$ for integers a and b, introducing primary numbers, and stating the analogue of the theorem given in section 3.4. The case $n = 3$ was dealt with by Jacobi and Eisenstein (and also in Gauss's manuscripts).

The search for higher reciprocity laws was the main motivation for Kummer in his construction of a theory of 'ideal numbers'. For further details on developments following Gauss, see [Smith, 1859–1865], [Dickson, 1919–1923] (§65) and [Neumann, 1980].

The material in Sections 1–4 of D.A. is currently the centre of much interest in cryptography (the technology of encoding and decoding information). Also in this context, further methods for factorizing large numbers as quickly as possible are being discussed and developed.

4 CONTENT AND RECEPTION OF SECTION 5

The fifth Section comprises more than half of D.A. and contains a systematic theory of the binary quadratic forms

$$f(x, y) = ax^2 + 2bxy + cy^2 = a^{-1}\left[(ax + by)^2 - (b^2 - ac)y^2\right] \quad (a, b, c \in \boldsymbol{Z}). \tag{10}$$

In the first place this involves a survey of Gauss's work from the beginning of his early studies in Göttingen in 1795 on the results of Fermat, Euler, Lagrange and Legendre. In the second place he succeeded in an incredibly short time in developing new methods which elevated the entire theory to a state of super-human perfection. The central problem for Fermat, Euler, Lagrange, Legendre and Gauss is the following question:

Representation problem. I. What values does $f(x, y)$ take when, with no loss of generality, the following general assumptions are made?

(i) $d(f) := b^2 - ac$, called the 'determinant' by Gauss in art. 154 and now known as the discriminant, is not a square;

(ii) the greatest common divisor $\gcd(a, 2b, c) = 1$, when f is called 'properly primitive' by Gauss in art. 226;

(iii) $x, y \in \boldsymbol{Z}$ with $\gcd(x, y) = 1$, in which case the equation $n = f(x, y)$ is called a 'proper representation' of n by f.

II. How can one describe all proper representations of a given number n by the form f (arts. 158, 180, 205)?

III. By which forms, that are somehow related to f, can a divisor of a number $n = f(x, y)$ (with $\gcd(x, y) = 1$) be represented?

Another important problem can be stated as follows.

Composition problem. Given properly primitive forms f and f', is there another properly primitive form F such that if $n = f(x, y)$, $n' = f'(x', y')$ are proper representations, then is there a proper representation $nn' = F(u, v)$?'

As far as part III of the representation problem is concerned, it was known to Lagrange and Legendre that every divisor m of $n = f(x, y)$ with $\gcd(n, d(f)) = 1$ has a representation $m = f'(u, v)$ by a quadratic form f' of the same discriminant as f, $d(f) = d(f')$ [Weil, 1984, ch. 4, sect. 4]. They solved Parts II and III using algorithms involving a fundamental understanding of forms with a given discriminant: equivalence and reduction of forms, finiteness of the class number, and the structure of solutions of Pell's equation ($t^2 - du^2 = 1, d > 0$).

The representation problem was in some sense solved definitively by Gauss. Firstly, he showed that in an equation $n = f(x, y)$ with the greatest common divisor $\gcd(x, y) = 1$ all prime factors are either divisors of $2d$ or belong to certain prime residue classes mod $4d$ (that depend only on d), and that when $\gcd(n, 2d) = 1$ the number n must lie in certain residue classes mod $4d$ that depend on d and f. (When $d < 0$, the assumption $\operatorname{sgn} n = \operatorname{sgn} a$ is added to those mentioned above.) Secondly, these necessary conditions are, in a weakened sense, also sufficient: they guarantee that when $\gcd(n, 2d) = 1$, n can be represented by a form of the same 'genus' as f, one of a set of forms of discriminant d determined by f. This result is demonstrably unimprovable by congruence conditions *alone*. The notion of the genus of a form (arts. 228–233) is one of Gauss's great innovations.

Lagrange and Legendre had also posed the composition problem (in a rather less precise form) and discussed it with some success in special cases [Weil, 1984, ch. 4]. Gauss settled the problem in a generality that left nothing to be desired. The composition problem has an affirmative answer if and only if $d(f)d(f')^{-1}$ is a rational square (arts. 234–244). For forms with fixed discriminant d, Gauss obtained a finite commutative group on a suitable partition of forms into classes (via the so-called proper equivalence, and only this!). This 'class group' (a later name) of d was the first example of a finite group not consisting of numbers or permutations.

The details of composition theory were notoriously difficult; they were simplified over the course of the ensuing decades. The first small contribution was made by the French translator of D.A. (remark to art. 235). Smith reported further significant simplifications in

[1859–1865, arts. 106–109], in the spirit of the theory of invariants; he developed in about 10 printed pages what had taken Gauss some 25 pages or so in D.A. (arts. 234–243).

On the Continent, the whole theory of composition was refounded by Dirichlet and Dedekind with the aid of irrational linear factors of forms; in other words, the path taken by Lagrange and Legendre was followed to the end. This path survives in the modern literature, in that it places Gauss's theory in the theory of quadratic number fields. Composition theory, like other theories of Gauss, shows clearly that his mathematics is the mathematics of explicit formulae and identities and the considerations of invariance and symmetry that support it.

5 CONTENT AND RECEPTION OF SECTION 7

The seventh Section concentrates on the equation $x^n = 1$, or $x^n - 1 = 0$, where n is a prime. Gauss set about solving this equation via a chain of auxiliary equations each of lowest degree (art. 342). This condition on the degree was new in the theory of equations.

Gauss first showed that the polynomial

$$\Phi_n(x) := (x^n - 1)(x - 1)^{-1} = x^{n-1} + x^{n-2} + \cdots + x + 1 \tag{11}$$

cannot be decomposed into factors with rational coefficients (art. 341); that is, in modern terminology, it is irreducible over $\boldsymbol{Q}$. He thus showed explicitly for the very first time that the polynomials of an infinite family are irreducible.

The equation $\Phi_n(x) = 0$ (where Gauss used X instead of $\Phi_n(x)$) has the property that its solutions, the so-called primitive nth roots of unity, consist precisely of the powers

$$\zeta, \zeta^2, \ldots, \zeta^{n-2}, \zeta^{n-1} \tag{12}$$

of an arbitrarily chosen solution ζ. This was already known before the time of Gauss. The novelty of his insight consisted of ordering the primitive nth roots of unity in a suitable way:

$$\zeta^{e(i)} \quad (0 \leqslant i \leqslant n-2) \text{ with } e(i) = g^i, \tag{13}$$

g being a primitive root mod n (compare (5)).

For any factorization $n - 1 = ef$, one can construct e new quantities $\eta_0, \eta_1, \ldots, \eta_{e-1}$, the so-called 'periods of f terms'. They satisfy an auxiliary equation of degree e with rational coefficients, while every primitive nth root ζ^k satisfies an equation of degree f whose coefficients depend rationally on the periods. By factorizing f further, one can obtain new auxiliary equations, and so on until the process terminates.

The solutions of $x^n = 1$ are represented by the vertices of a regular n-gon in the complex plane. A geometrically interesting case arises when n is a prime of the form $2^m + 1$ (a Fermat prime), such as $n = 17 = 2^4 + 1$. Then all the auxiliary equations have degree 2, and their solution corresponds to ruler-and-compass constructions. It follows that, for example, the regular 17-gon is constructible by ruler and compasses. For the 19-year-old Gauss, this sensational advance in a problem of more than 2000 years standing came at the

beginning of the theory of cyclotomy, and convinced him to devote himself exclusively to mathematics.

Gauss further showed (with some gaps that he specified) that every equation $x^n - 1 = 0$ (where n is prime) can be solved via a chain of pure, or binomial, equations $y^m - a = 0$ of lowest possible degree. Prior to Gauss, only the case $n = 11$ had been settled, by A.-T. Vandermonde (1735–1796) in his paper [1774]. We are thus concerned with the problem, of paramount interest up to the middle of the 19th century, of solving 'algebraic' equations by radicals, or, more precisely (as was apparently first recognized by Gauss), by irreducible radicals.

S.F. Lacroix (1765–1843) commented on the theory of cyclotomy in the third edition of his *Compléments des éléments d'algèbre* (1803). In Kazan, the young Lobachevsky had occupied himself with D.A. since 1811 under the guidance of M. Bartels (1769–1836), and in 1813 he wrote a piece 'on the solution of the algebraic equation $x^n - 1 = 0$'.

For N.H. Abel (1802–1829), the seventh Section of the D.A. formed the definitive model for a theory of soluble equations (compare §29). Galois also oriented himself around the 'method of Mr. Gauss'. Later, Dedekind had this to say about number theory [1873, 410]:

> it soon becomes clear that [...] the cyclotomy form an inexhaustable source of ever newer and more significant advances in number theory. One can say that almost all subsequent progress [...] either owes its inception directly to cyclotomy or, which is in some cases even more remarkable, arises in a previously unsuspected connection with cyclotomy.

As already mentioned, Gauss had in fact discovered that every square root is an integer linear combination of roots of unity. Thus we can say with a grain of salt that Section 7 contains Sections 4 and 5.

6 FAME AND REACTIONS

D.A. quickly gained a reputation among specialists and made Gauss famous. Reaction was first heard in France, then in Germany and Russia. Lagrange wrote as follows to Gauss in 1804: 'Your Disquisitiones immediately places you among the first rank of geometers' (*Oeuvres*, vol. 14 (1892), 298–299). The *Bureau des Longitudes* in Paris ordered 50 copies [Küssner, 1979, 119], and the French translation appeared as early as 1807.

The various parts of D.A. were received in different ways. The correspondence between Gauss and Sophie Germain (1776–1831) (in Gauss *Works*, vol. 10, pt. 1, 70–74) showed that the latter had penetrated so deeply into the book that she was able to investigate independently questions on higher-power residues. Legendre reacted in the second (1808) and third (1830) editions of his *Essai*. A part of the theory of cyclotomy appears in the second edition (arts. 484ff.) along with applications to number theory. On the other hand, Legendre gave no comprehensive treatment of the results in D.A. on quadratic forms, because Gauss's methods were so specialized that he could only have included them as part of a wide detour or as a bald translation of Gauss's work.

In the first 25 years after its appearance, D.A. essentially exercised only a gradual influence, and the earliest applications did not visibly go beyond the work of Gauss in either

method or substance. Further developments in depth were first pursued by Abel, Jacobi, Dirichlet, Galois, A.-L. Cauchy (1789–1857), C. Hermite (1822–1901) and Eisenstein, and later by Kummer, Kronecker and Smith as representatives of a new generation. Kummer mentioned in his speech in commemoration of Dirichlet that the latter kept a well-thumbed copy of D.A. permanently on his desk. Kummer also said 'that in more than 20 years after it appeared, none of the mathematicians alive at the time mastered it completely [...] Dirichlet was the first not only to understand it completely but also to make it accessible to others' [1860, 316]. Eisenstein also got intensively to grips with D.A.

Gauss received a noteworthy response from outside the circle of specialists: namely, the influential philosopher G.W.F. Hegel (1770–1831), who had a copy of D.A. in his private library [Mense, 1993] along with a paper on number theory by Gauss. Hegel wrote as follows in his *Wissenschaft der Logik* (*The science of logic*) in 1814: 'Thus the solution of the equation $x^m - 1 = 0$ using the sine, as well as the implicit algebraic solution found by Gauss by considering the residue of $x^{m-1} - 1$ divided by m and the so-called primitive roots, which form one of the most important extensions of analysis in recent times, is a synthetic solution, in that the auxiliary notions of *sine* and residue are not inherent in the problem itself' [Hegel, 1949, 286]. To explain, m is a prime (art. 350 of D.A.), 'residue' means residue class mod m, and the solutions of the congruence $x^{m-1} - 1 \equiv 0 \pmod{m}$ are precisely the prime residue classes mod m, on whose structure the investigation of the primitive mth roots of unity depends.

Nobody who speaks of the number theory and algebra of the last 200 years can remain silent about their sources in the *Disquisitiones arithmeticae* of Gauss. Notwithstanding all the new proofs of results in detail, this work belongs to the 'eternal canon' of mathematics, and thus of human culture.

BIBLIOGRAPHY

Biermann, K.-R. 1977. 'Aus unveröffentlichten Aufzeichnungen des jungen Gauss (zum 200. Geburtstag von C.F. Gauss)', *Wissenschaftliche Zeitschrift der Technischen Hochschule Ilmenau, 23*, no. 4, 7–24.

Biermann, K.-R. 1990. *Carl Friedrich Gauß. Der "Fürst der Mathematiker" in Briefen und Gesprächen*, München: C.H. Beck.

Dedekind, R. 1871. 'Ueber die Composition der binären quadratischen Formen', in J.P.G. Lejeune Dirichlet, *Vorlesungen über Zahlentheorie*, 2nd ed., Braunschweig: Vieweg, 380–497.

Dedekind, R. 1873. Notice of P. Bachmann, *Die Lehre von der Kreisteilung und ihre Beziehungen zur Zahlentheorie*, in *Literaturzeitung der Zeitschrift für Mathematik und Physik, 18*, 14–24. [Repr. in *Gesammelte mathematische Werke*, vol. 3, 1932, 408–419: cited here.]

Dickson, L.E. 1919, 1920, 1923. *History of the theory of numbers*, 3 vols., Washington, DC: Carnegie Institute. [Repr. New York, Chelsea, 1992. See §65.]

Gauss, C.F. *Works. Werke*, 12 vols., Berlin and Leipzig: Teubner, 1863–1933. [Repr. Hildesheim: Olms, 1973.]

Gauss, C.F. Diary. Manuscript, published in 'Abdruck des Gaußschen Tagebuchs (Notizenjournal) mit Erläuterungen', in *Works*, vol. 10, pt. 1 (1917), 483–574. [Annotated French trans. by P. Eymard and J.-P. Lafon, in *Revue d'histoire des sciences, 9* (1956), 21–51. German trans. by E. Schuhmann in *Mathematisches Tagebuch 1796–1814* (ed. various), Leipzig: Geest und Portig, 1979. Eng. trans. by J.J. Gray in *Expositiones mathematicae, 2* (1984), 97–130.]

Gauss, C.F. 1825, 1831. 'Theoria residuorum biquadratorum', *Commentarii Societatis Regiae Göttingensis, 6*, 27–56; *7*, 89–148. [Repr. in *Works*, vol. 2, 65–178.]

Goldstein, C. et alii 2001. *Bibliography of secondary literature on the history of number theory after 1800*, Université de Paris-Sud, Mathématiques, Bâtiment 425, 91405 Orsay, France, Prépublications n°33, 18 pp.

Goldstein, C., Schappacher, N. and Schwermer, J. (eds.) In preparation. *The shaping of arithmetic: number theory after Carl Friedrich Gauss's "Disquisitiones Arithmeticae"*, Berlin: Springer.

Hegel, G.W.F. 1949. *Wissenschaft der Logik*, 1st ed. 1814. [Edition used publ. 1949.]

Klein, F. 1926. *Vorlesungen über die Entwicklung der Mathematik im 19. Jahrhundert*, vol. 1, Berlin: J. Springer. [Repr. New York: Chelsea, 1967.]

Kummer, E.E. 1860. 'Gedächtnisrede auf Gustav Peter Lejeune-Dirichlet', *Abhandlungen der Königlichen Akademie der Wissenschaften zu Berlin*, 1–36. [Repr. in *Collected papers*, vol. 2, Berlin: Springer, 1975, 721–756. Also in Dirichlet, *Mathematische Werke*, vol. 2, 1889 (repr. New York: Chelsea, 1969), 309–344 (cited here).]

Küssner, M. 1979. *Carl Friedrich Gauß und seine Welt der Bücher*, Göttingen: Musterschmidt.

Legendre, A.-M. 1798. *Essai sur la théorie des nombres*, 1st ed., Paris: Duprat. [Later eds. 1808 and 1830.]

Maennchen, P. 1918. 'Gauss als Zahlenrechner', repr. as (Gauss *Works*), vol. 10, pt. 2, no. 6.

May, K.O. 1972: 'Gauss, Carl Friedrich', in *Dictionary of scientific biography*, vol. 5, New York: Scribners, 298–315. [Repr. in *A biographical dictionary of mathematicians*, vol. 2, New York: Scribners, 1991, 860–877.]

Mense, A. 1993. 'Hegel's library: the works on mathematics, mechanics, optics and chemistry', in M.J. Petry (ed.), *Hegel and Newtonianism*, Dordrecht: Kluwer, 669–709.

Merzbach, U. 1981. 'An early version of Gauss' *Disquisitiones Arithmeticae* ', in J.W. Dauben (ed.), *Mathematical perspectives. Essays on mathematics in its historical development*, New York: Academic Press, 167–178.

Merzbach, U. 1984. *Carl Friedrich Gauss. A bibliography*, Wilmington, Delaware: Scholarly Resources Inc.

Milnor, J. 1971. *Introduction to algebraic K-theory*, Princeton: Princeton University Press.

Neumann, O. 1980. 'Zur Genesis der algebraischen Zahlentheorie', *NTM Schriftenreihe, 17*, no. 1, 32–48; no. 2, 38–58.

Smith, H.J.S. 1859–1865. 'Report on the theory of numbers', parts 1–6, in *Report of the British Association for the Advancement of Science* for those years. [Repr. in *The collected mathematical papers*, vol. 1, Oxford: Clarendon Press, 1894 (repr. New York: Chelsea, 1965), 38–364.]

Vandermonde, A.-T. 1774. 'Mémoire sur la résolution des équations', *Histoire et mémoires de l'Académie des Sciences de Paris*, (1771), 365–416. [German trans. in *Abhandlungen aus der reinen Mathematik* (ed. C. Itzigsohn), Berlin: 1888, 1–64.]

Weil, A. 1984. *Number theory. An approach through history from Hammurapi to Legendre*, Boston: Birkhäuser. [German trans. by H. Pieper: *Zahlentheorie. Ein Gang durch die Geschichte von Hammurapi bis Legendre*, Basel: Birkhäuser, 1992.]

CHAPTER 23

CARL FRIEDRICH GAUSS, BOOK ON CELESTIAL MECHANICS (1809)

Curtis Wilson

In this work Gauss offered new methods of determining the orbital parameters of planetary motion. They were more compact than those of Lagrange and Laplace, and established more careful ways of treating questions of precision and observational error.

First publication. Theoria motus corporum coelestium in sectionibus conicis solem ambientium, Hamburg: F. Perthes and I.H. Besser, 1809. xii + 227 pages, 21 tables and figures.

Later edition. In Gauss *Werke*, vol. 7, Leipzig: Teubner, 1871, 1–282. [Repr. 1905.]

English translation. Theory of the motion of the heavenly bodies moving about the Sun in conic sections (trans. C.H. Davis), Boston: Little Brown, 1857. [Repr. New York: Dover, 1963.]

Russian translation. Teorija dwishenija nebesnych tel, obrastschich wokrug Solnca po konitscheskim setsschenijam (trans. Dogel), Moscow: 1861.

French translation. Théorie du mouvement des corps célestes parcourant des sections coniques autour du soleil (trans. E. Dubois), Paris: A. Bertrand, 1864.

German translation. Theorie der Bewegung der Himmelskörper welche in Kegelschnitten die Sonne umlaufen (trans. C. Haase), Hannover: C. Meyer, 1865.

Related articles: Newton (§5), Lagrange on mechanics (§19), Laplace (§18, §24).

1 BIOGRAPHICAL SKETCH

Born into a poor family in Braunschweig, Carl Friedrich Gauss (1777–1855) was supported in his education and mathematical work till 1807 by ducal stipends. He attended the University of Göttingen from October 1795 to September 1798. In 1796 he discovered the

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

constructibility of the 17-sided regular polygon. His doctoral dissertation (1799) gave the first proof of the fundamental theorem of algebra. His *Disquisitiones arithmeticae* (1801) would be fundamental for later number theory (§22). His work on the determination of orbits began in September 1801. In 1807 he accepted a professorship at Göttingen University (as director of the Observatory), the position he would occupy till his death.

Among Gauss's publications after the *Theoria motus* were his memoir on the hypergeometric series (1812), new demonstrations of the fundamental theorem of algebra (1815, 1816, 1849), a memoir on determining perturbations by considering the perturbing body's mass as distributed round its orbit (1818), a new interpretation of the method of least squares (1823), a memoir on the curvature of surfaces (1828), memoirs on the measure of the Earth's magnetic field in absolute units (1832 and later), and investigations in geodesy (1842, 1846, 1847). Many of his papers report astronomical observations, or discuss observational methodology.

2 INCEPTION OF GAUSS'S WORK IN ORBIT DETERMINATION

On 1 January 1801, Giuseppe Piazzi in Palermo discovered a comet or planet in the constellation of Taurus, detectable only telescopically. He observed it through 11 February, when illness interrupted his observations. He informed three astronomers of his discovery, and in May sent his detailed observations to J.J. Lalande in Paris, asking that publication be postponed.

Since the 1770s two astronomers, J.E. Bode of Berlin and Franz Xaver von Zach (1754–1832) of Gotha, had entertained the notion of a missing planet between Mars and Jupiter. A numerical series due to J.D. Titius, publicized by Bode in 1772, gave approximate mean solar distances of the known planets, but predicted a planet in this 'gap'. It received surprising corroboration in 1781 with the discovery of Uranus, a planet whose nearly circular orbit had a radius close to the next term after Saturn in the series. In autumn 1800 Zach and other German astronomers formed a society to promote systematic search for the missing planet.

In spring 1801 the question arose: might Piazzi's 'comet' be the quarry sought? It must be re-discovered! From June onward, Zach's monthly reports in a periodical which he published, the *Monatliche Correspondenz zur Beförderung der Erd- und Himmels-Kunde* (hereafter, '*MC*') gave an ongoing account of the search.

The July issue reported the efforts of J.C. Burckhardt, in Paris, to put an orbit to Piazzi's observations. Parabolic orbits, Burckhardt found, were unsatisfactory; circular orbits could accommodate more of the data. He proposed an approximate elliptical orbit, but in agreement with P.S. Laplace (1749–1827), held that an accurate orbit determination would require more observations.

Through late summer and autumn, cloudy weather prevented a systematic search. In the September issue, Zach published Piazzi's revised observations. Gauss, a subscriber to the *MC*, set about determining an orbit.

The November issue of the *MC* contained a review of Piazzi's memoir on his discovery. Finding parabolic trajectories hopeless, he had derived two circular orbits with radii $2 \cdot 7067$ and $2 \cdot 68626$ astronomical units. From the second of these Zach computed an ephemeris

for November and December. Piazzi named the planet *Ceres Ferdinandea*, thus honoring Sicily's ruler.

Zach now received Gauss's results, and to them devoted his entire report in the December issue. Gauss had computed four different elliptical orbits, each based on a different trio of observations; the four sets of elements were in near agreement with each other and with the 19 observations Piazzi had considered undoubtful. (For the second set of elements, the root mean square errors in longitude and latitude are $7'' \cdot 33$ and $4'' \cdot 35$; for the third set of elements they are $5'' \cdot 86$ and $2'' \cdot 60$.) Gauss put the planet in January 1801 about a quadrant past aphelion; and assigned it a considerably higher eccentricity than had Burckhardt, so that in December 1801 the planet would be 6° or 7° farther east than any of the other proposed orbits implied. He gave positions for Ceres at 6-day intervals from 25 November to 31 December.

The weather continued unpropitious. As Zach reported in the January 1802 issue of the *MC*, in the early morning hours of 7–8 December he clocked a star very close to Gauss's prediction for Ceres, but bad weather on the following nights prevented verification.

As he reported in the February 1802 issue, early on 1 January Zach discovered the planet some 6° east of its December position, and through January he followed its motion, which agreed closely with Gauss's orbital elements. Wilhelm Olbers (1758–1840) also re-discovered the planet, reporting the fact to the newspapers, where Gauss read about it. Gauss's ellipse, exclaimed Zach, was astonishingly exact. 'Without the ingenious efforts and calculations of Dr. Gauss, we should probably not have found Ceres again; the greater and more beautiful part of the achievement belongs to him'.

3 GAUSS'S EARLY WORK ON ORBIT-DETERMINATION

Of Gauss's earliest orbit-computations, only sparse indications remain. After the rediscovery of Ceres, and one by Olbers in March 1802 of another planet, Pallas, for which Gauss also computed an orbit, he wrote an account of his procedures [Gauss, 1809], which he sent to Olbers in August 1802. We review these procedures, for comparison with the mature methods of the *Theoria motus* (hereafter, '*TM*').

As Gauss informed Olbers in an accompanying letter, the surmise leading to his new method had come to him five years earlier, on first reading Olbers's *Abhandlung über die leichteste und bequemste Methode die Bahn eines Cometen aus einigen Beobachtungen zu berechnen* (1797). Gauss's first steps were Olbersian; in a general way his procedures would remain so.

Olbers, a well-known comet-finder, had searched the literature for a convenient method of determining cometary orbits. The direct algebraic routes proposed by J.L. Lagrange and Laplace led to seventh-degree equations and other algebraic complications. Laplace's method, in A.-G. Pingré's view the best (*Cométographie*, 1784), required an initial derivation of a mean geocentric position with its first and second time-derivatives; Olbers questioned its expediency. In fact, small, hardly avoidable errors in the derivatives could lead to large errors in the computed results.

With borrowings from predecessors, Olbers devised a simpler procedure. The first step was to obtain an approximate solution, using linear or quadratic equations, some rigorous,

the others nearly so. Secondly, the approximate solution was to be refined. Gauss adopted this two-stage approach. The focus was not so much on rigor in formulas as on the delicate fitting of computed elements to data.

Both Gauss and Olbers started by assuming that the orbit of the comet or planet lay in a plane passing through the Sun's center S, and that the plane of the Earth's orbit also contained S. Three observed positions (P, P', P'') of the orbiting body were taken as data. Let their heliocentric rectangular coordinates be (x, y, z), (x', y', z'), (x'', y'', z''), and let twice the triangular areas PSP', $P'SP''$, $P''SP$, be

$$n'' = rr' \sin(v' - v), \qquad n = r'r'' \sin(v'' - v'), \qquad n' = r''r \sin(v'' - v), \tag{1}$$

where r, r', r'' are the three *radii vectores*, and v, v', v'' the three longitudes in orbit. (To facilitate comparisons, here and later we substitute the symbols of *TM* for those of the summary [Gauss, 1809].) The assumptions imply that

$$0 = nx - n'x' + n''x'', \qquad 0 = ny - n'y' + n''y'', \qquad 0 = nz - n'z' + n''z''. \tag{2}$$

Analogous propositions hold for the projections of the areas n, n', n'' onto the three coordinate planes.

At this point Olbers, followed by Gauss, introduced the further assumption that the radius vector from Sun to heavenly body at the time τ' of the second observation cuts the chord PP'' in the ratio of the times $(\tau'' - \tau') : (\tau' - \tau)$. The analogous proportion is assumed for the Earth. This means that the triangular areas n, n'' (and also $\Delta ESE'$, $\Delta E'SE''$, where E, E', E'' are the three positions of the Earth) are proportional to the time-differences, whereas by Kepler's areal rule it is the corresponding *sectors* that are thus proportional. As Olbers pointed out, the approximation is best when τ' falls midway between τ and τ''. On the foregoing assumption Gauss obtained expressions for δ and δ'', the projections of EP and $E''P''$ onto a plane through the Earth's center parallel to the ecliptic, in terms of δ', the projection of $E'P'$ onto the same plane:

$$\delta = \frac{\tan\beta' \cdot \sin(\alpha'' - L') - \tan\beta'' \sin(\alpha' - L')}{\tan\beta \cdot \sin(\alpha'' - L') - \tan\beta'' \sin(\alpha - L')} \cdot \frac{\tau'' - \tau}{\tau'' - \tau'} \cdot \delta', \tag{3}$$

$$\delta'' = \frac{\tan\beta \cdot \sin(\alpha' - L') - \tan\beta' \sin(\alpha - L')}{\tan\beta \cdot \sin(\alpha'' - L') - \tan\beta'' \sin(\alpha - L')} \cdot \frac{\tau'' - \tau}{\tau' - \tau} \cdot \delta', \tag{4}$$

where β, β', β'' are the geocentric latitudes, $\alpha, \alpha', \alpha''$ the geocentric longitudes of the body in the three observations, and L' the heliocentric longitude of the Earth in the second observation. Olbers obtained a single equation, the quotient (3)–(4).

Olbers used his equation to compute, for an arbitrarily chosen value of δ, the value of δ'' given by (3)–(4). Since

$$\delta = EP\cos\beta \quad \text{and} \quad \delta'' = E''P''\cos\beta'', \tag{5}$$

the corresponding values of EP and $E''P''$ can then be found. Then, since SE and SE'' are known from solar theory, and the angles SEP and $SE''P''$ from observation and solar theory,

triangles PSE and $P''SE''$ can be solved, and the *radii vectores* $r = SP$, $r' = SP'$, $r'' = SP''$, together with the chord PP'', computed.

At this point Olbers invoked a relation in parabolic motion that Leonhard Euler had established in 1743:

$$(r + r'' + PP'')^{3/2} - (r + r'' - PP'')^{3/2} = 6\kappa(\tau'' - \tau), \tag{6}$$

where κ is a constant such that time measured as $\kappa\tau$ adds up to 2π in a sidereal year. This relation enabled Olbers to compute the time interval $(\tau'' - \tau)$ implied by his choice of δ. With a second choice of δ, he repeated the calculation; and then, by interpolation, obtained values of δ, δ'' agreeing with the observed time interval.

Although the Eulerian relation had been generalized to all conic sections by Lambert, Gauss did not employ it. He applied the standard formula for a conic to the positions P, P', P'':

$$\frac{1}{r} = \frac{1}{p}\big[1 - e\cos(v - \pi)\big], \quad \frac{1}{r'} = \frac{1}{p}\big[1 - e\cos(v' - \pi)\big], \quad \frac{1}{r''} = \frac{1}{p}\big[1 - e\cos(v'' - \pi)\big], \tag{7}$$

where p is the semi-parameter, e the eccentricity, and π longitude of the aphelion. Multiplying these three equations by $\sin(v'' - v')$, $\sin(v - v'')$, $\sin(v' - v)$ respectively and adding, he obtained, after some transformations,

$$\frac{n - n' + n''}{-n'} = -\frac{2r'}{p} \cdot \frac{\sin 1/2(v'' - v')\sin 1/2(v' - v)}{\cos 1/2(v'' - v)}. \tag{8}$$

To substitute for p, Gauss used the known theorem:

$$\frac{\Delta(area)}{\Delta m} = \frac{a^{3/2} \cdot \sqrt{p}}{2}, \tag{9}$$

where $\Delta(area)$ is the area swept out by the radius vector, Δm is the concomitant change in mean anomaly, and a is the semi-major axis (for an elliptical or hyperbolic orbit). From (8) it can be deduced that

$$p = \frac{4gg''}{a^3(m' - m)(m'' - m')} = \frac{4gg''}{A^3(M' - M)(M'' - M')}, \tag{10}$$

where $g'' = \text{sector}\, SPP'$; $g = \text{sector}\, SP'P''$; m, m', m'' are the mean anomalies of the body in its three positions; A is the semi-major axis of the Earth's orbit (usually taken as the unit); and M, M', M'' are the three mean anomalies of the Earth in its three positions. The equality of the two denominators follows from Kepler's third law. From (8) and (10), Gauss obtained the approximate relation

$$\frac{R'}{\delta'}\left(\frac{1}{R'^3} - \frac{1}{r'^3}\right) = \frac{2}{A^3(M' - M)(M'' - M')} \cdot \frac{[\pi\pi'\pi'']}{[\pi E\pi'']}, \tag{11}$$

where R', r' are the *radii vectores* of the Earth and of the celestial body in the middle observation, and

$$\frac{[\pi\pi'\pi'']}{[\pi E'\pi'']} = \frac{\tan\beta'\sin(\alpha''-\alpha) - \tan\beta\cdot\sin(\alpha''-\alpha') - \tan\beta''\sin(\alpha'-\alpha)}{\tan\beta\cdot\sin(L'-\alpha'') - \tan\beta''\sin(L'-\alpha)}. \tag{12}$$

In his letter to Olbers, Gauss called equation (11) 'the most important part of the whole method and its first foundation'. If the time-differences $\tau'-\tau$, $\tau''-\tau'$, $\tau''-\tau$, as well as n, n', n'', are viewed as infinitely small quantities of the first order, he stated that (11) will be correct down to quantities of the second order of smallness, provided that τ' is midway between τ and τ'', otherwise to quantities of the first order. Every quantity on the right-hand side of (11) is known from observation or solar theory. Using (11) along with the formula

$$\frac{R'/\delta'}{R'/r'} = \sqrt{1+\tan^2\beta + \frac{R'^2}{\delta'^2} + 2\frac{R'}{\delta'}\cos(\lambda'-L')}, \tag{13}$$

which is a near approximation to the cosine law applied to $\Delta E'SP'$, accurate to the second order of smallness. Gauss proceeded by a 'cut and try' process with rapid convergence to values for r' and δ' satisfying both (11) and (13).

Given satisfactory values for r' and δ', (3) and (4) can be used to find the corresponding values of δ and δ'', and thence r and r'' . What remains is to determine the orbital elements. The coordinates of P, P', P'', now known, together with the Sun's position $(0,0,0)$, lead straightforwardly to values for Ω, the longitude of the ascending node, and i, the orbit's inclination; and the values of (r,v), (r',v'), (r'',v'') substituted into equations (7) yield values for p, π, e. Gauss preferred, however, to obtain p from the equation

$$\int_v^{v''} r^2\,dv = A^{3/2}(M''-M)\sqrt{p}, \tag{14}$$

using approximations due to Roger Cotes to evaluate the integral on the left. He then obtained π and e from the first and third of equations (7), and with the values so obtained computed the middle observation to provide a check on the entire computation.

In his 'Summarische Übersicht' of the book [Gauss, 1809] described several methods of refining the orbital elements initially found. For instance, δ and δ'' could each be altered by a small amount, the elements re-determined in each case, and the middle observation re-computed. Interpolation would then lead to elements giving a more accurate value of the middle observation.

More than anyone earlier, Gauss was focussing on close fitting of computed results to observations. In *TM* he will introduce new procedures, eliminating reliance on (2) and (3) and the 'cut and try' procedure used in resolving (11). His focus on fitting results to observations will remain.

4 THE DETERMINATION OF ORBITS IN *TM*

The contents of Gauss's book are summarized in Table 1; on its background and genesis see [Reich, 1998, 2001]. Reserving our review of Book I till later, we turn to Section 1

Table 1. Contents by Sections of Gauss's book. xii + 227 pages. The titles are translated.

Sect.; arts.	**Title**
	Preface.
Book I	*General relations between those quantities by which the motions of the heavenly bodies about the Sun are defined.*
1; 1–46	Relations pertaining simply to position in orbit.
2; 47–77	Relations pertaining simply to position in space.
3; 78–109	Relations between several places in orbit.
4; 110–114	Relations between several places in space.
Book II	*Investigation of the orbits of heavenly bodies from geocentric observations.*
1; 115–163	Determination of an orbit from three complete observations.
2; 164–171	Determination of an orbit from four observations, of which two only are complete.
3; 172–189	The determination of an orbit satisfying as nearly as possible any number of observations whatever.
4; 190–192	On the determination of orbits, taking into account the perturbations.

of Book II, which treats the problem that Gauss calls 'the most important in this work'—the determination of an orbit wholly unknown, starting from three complete observations. (An observation is complete if it furnishes two coordinates specifying the body's place on the celestial sphere at a given time: its longitude and latitude, or its right ascension and declination.) In the case of orbits nearly coinciding with the ecliptic—dealt with in Section 2 of Book II—Gauss derives the orbit from four observations, two complete and two of longitude merely.

In both cases, the initial problem is reduced to the solution of two equations, $X = 0$, $Y = 0$, in two unknowns, x and y (art. 119). The latter need not be orbital elements, but must be so connected with the elements as to permit their deduction. Nor need X and Y be explicit functions of x and y, but they must be so connected with x, y that, from given values of x, y, the functions X, Y can be computed. Gauss's principal concern (art. 120) is that x, y be so selected, and X, Y so arranged, that X, Y may depend in the simplest manner on x, y, and that the elements may follow easily from x, y. A further concern is how values of x, y closely satisfying the equations may be had without excessive labor. From such approximations, values of x, y having all needful accuracy can generally be got by linear interpolation.

For determining an orbit quite unknown, the observations should be fairly close together: the accuracy of the approximations increases when the heliocentric motion between observations is less. But then the influence of observational error increases; hence a compromise must be struck. The 47-day spread among Piazzi's observations of Ceres, encompassing only 3° of heliocentric motion, proved quite satisfactory for determining the orbit.

Leading to Gauss's new method were the following considerations. In art. 114 he shows that, if the mutual ratios of the three double areas n, n', n'' were known, then the exact values of $\delta, \delta', \delta''$ would be given algebraically, without 'cut and try' procedures. Thus it is

exactly true that

$$a\delta' = b + c\frac{n}{n'} + d\frac{n''}{n'}. \tag{15}$$

Here $a = (0.\text{I}.2)$, $b = -(0.\text{I}.2)D'$, $c = (0.\text{O}.2)D$, $d = (0.\text{II}.2)D''$, where the symbol (0.1.2) has the same meaning as $[\pi\pi'\pi'']$ in (11) above; the expressions symbolized by (0.O.2), (0.I.2), and (0.II.2) are derived through replacement of α', β' in (0.1.2) by the Earth's heliocentric longitude and latitude in the first, second, and third observations, respectively; and D, D', D'' are the Earth–Sun distances in the same three observations.

To determine δ' by (14), we must substitute approximations for the ratios $n : n', n'' : n'$. Suppose we use the ratios of the time intervals for this purpose (art. 131). Gauss lets $\theta, \theta', \theta''$ stand for $k(\tau'' - \tau'), k(\tau'' - \tau), k(\tau' - \tau)$, respectively, where k is a constant for all bodies orbiting the Sun (art. 1); thus $\theta, \theta', \theta''$ are as the sectors $P'SP'', PSP'', PSP'$. Next he defines η, η', η'' so that $n\eta = \theta, n'\eta' = \theta', n''\eta'' = \theta''$. If n, n', n'' are regarded as quantities of the first order of smallness, Gauss tells us that $\eta - 1, \eta' - 1, \eta'' - 1$ will, generally speaking, be quantities of the second order; therefore $\theta/\theta', \theta''/\theta'$ will differ from $n/n', n''/n'$ by quantities of the second order. Nevertheless, he finds the proposed substitution 'wholly unsuitable'.

In the expression for δ' drawn from (14), each term will have as denominator the quantity $a = (0.1.2)$, which is of the third order of smallness. In the numerators $c = (0.\text{O}.2)D$ and $d = (0.\text{II}.2)D''$ are of the first order. Hence an error of the second order in the substitutions for $n/n', n''/n'$ produces an error of order zero in the values of δ': the error can be larger than the quantities sought.

Much of the error comes from assuming n' proportional to θ' while assuming n and n'' proportional to θ and $\theta'' : \eta'$ is distinctly larger than either η or η''. For, as the orbit is ever concave toward the Sun, the whole sector PSP'' bears a larger ratio to its triangle PSP'' than do the component sectors $PSP', P'SP''$ to their triangles. Let (15) be written in the form

$$a\delta' = b + \frac{cn + dn''}{n + n''} \cdot \frac{n + n''}{n'}, \tag{16}$$

where n' occurs only in the final factor. It can be shown that

$$\frac{n + n''}{n'} = 1 + \frac{\theta\theta''}{2\eta\eta'' rr'r'' \cos 1/2(v'' - v') \cos 1/2(v'' - v) \cos 1/2(v' - v)}. \tag{17}$$

By contrast, the proposed substitution, $(\theta + \theta'')/\theta'$, is equal to 1.

As a remedy, Gauss takes for x, y the quantities

$$P = \frac{n''}{n}, \qquad Q = 2\left(\frac{n + n''}{n'} - 1\right)r'^3, \tag{18}$$

and rewrites (16) as

$$a\delta' = b + \frac{c + dP}{1 + P}\left(1 + \frac{Q}{2r'^3}\right). \tag{19}$$

If now for P, Q we substitute the approximations $P = \theta''/\theta$, $Q = \theta\theta''$, the parenthesis on the right of (19) will err by an error of only the fourth order. If the radius vector SP' divides the chord PP'' in the middle, the value of δ' will differ from its correct value by an error of the second order, otherwise by an error of the first order. Gauss's new procedure, based on the considerations just explained, is described and illustrated in arts. 136–171.

5 THEMES AND TOPICS OF BOOK I

TM begins with a statement of the planetary laws of 'our own Kepler', but in a Newtonian form applicable to all conic sections, and viewed as a consequence of the Sun's gravitational attraction. In art. 1 Gauss introduces a constant k for all bodies orbiting the Sun: it is an expression of Kepler's third law:

$$k = \frac{g}{t\sqrt{p(1+\mu)}} = 0.01720209895, \tag{20}$$

where g is twice the area swept out by the radius vector, p is the parameter of the conic, and $1 + \mu$ the sum of the masses of the Sun $(= 1)$ and orbiting body. Gauss's value of k, based on the orbital constants of the Earth accepted in his day, is retained today to define the astronomical unit of distance.

Section 1 concerns 'relations pertaining to position in orbit'; with separate treatments for the ellipse (arts. 6–17), parabola (arts. 18–20), and hyperbola (arts. 21–29). We restrict our comments to the elliptical case.

Following a policy adopted about 1800 by the French *Bureau des Longitudes*, Gauss in *TM* measures anomalies in elliptical orbits from perihelion, as is necessarily done in parabolic and hyperbolic orbits. 'Kepler's equation' takes the form

$$M = E - e\sin E, \tag{21}$$

with a minus sign on the right rather than the plus sign used by Kepler. M is the mean anomaly, expressible as $kt/a^{3/2}$ if as in most cases the mass μ of the orbiting body can be neglected. The eccentric anomaly, linking M and the true anomaly v, is defined by $\tan\frac{1}{2}E = \sqrt{\frac{1-e}{1+e}}\tan\frac{1}{2}v$. The radius vector given in terms of it is $r = a(1 - e\cos E)$, more convenient for integrations or differentiations than the formula in terms of the true anomaly (compare (7) above).

'Kepler's problem' refers to the determination of E for a given value of M; direct procedures are unavailable, (21) being transcendental. Rather than using a series expansion as does [Laplace, 1799, Bk. II, ch. 3, art. 22], Gauss recommends solving the equation $E = M + e\sin E$ by trial (art. 11). Suppose ε nearly enough approximates E to permit refinement by linear interpolation. Let

$$\frac{\partial \log\sin\varepsilon}{\partial\varepsilon} = \lambda, \qquad \frac{\partial \log e\sin\varepsilon}{\partial e\sin\varepsilon} = \mu. \tag{22}$$

Then a more correct value of E will be $\varepsilon + x$, where

$$x = \frac{\mu}{\mu+\lambda}(M + e\sin\varepsilon - \varepsilon). \tag{23}$$

Here as throughout *TM* Gauss assumes use of 7-place logarithms to base 10; in most calculations these yield results precise to $0 \cdot 1$ arcsecond, equal or superior to the observational precision attainable in the early 1800s.

Gauss is the first to supply, in arts. 30–32 of *TM*, rules governing the propagation of error in computations. Numbers taken from logarithmic or trigonometric tables, he points out, are liable to an error amounting to half a unit in the last figure; e.g., 0.00000005 in 7-place tables. When such approximate quantities are combined by addition or subtraction, the possible errors add. In multiplication or division, the maximum error is increased or diminished in the same ratio as the quantity itself. Suppose, for instance, that the true anomaly v is to be computed from the eccentric anomaly E by

$$\log\tan\frac{v}{2} = \log\tan\frac{E}{2} + \log\sqrt{\frac{1+e}{1-e}}. \tag{24}$$

If the maximum error in taking a logarithm or antilogarithm from the tables is ω, the possible error in computing $\log\tan(v/2)$ will be 2ω, and in obtaining $v/2$ will be

$$\frac{3\omega\partial(v/2)}{\partial\log\tan(v/2)} = \frac{3\omega\sin v}{2\lambda}, \tag{25}$$

where λ is the modulus of the logarithms used. The maximum error in v will be twice as great. With 7-place logarithms to base 10, this error is $0'' \cdot 0712$.

When e is close to 1, obtaining precise enough values of the true anomaly from the time or *vice versa* requires special methods. Gauss provides these in arts. 33–46, with an auxiliary table to facilitate their application.

Section 2 treats 'relations pertaining simply to position in space'. The position of the orbit must be specified with respect to the orbital plane, and the position of the orbital plane with respect to a coordinate system, ecliptic or equatorial. Gauss introduces the practice of giving the orbital longitude of the perihelion as the sum of the longitude of the ascending node (where the planet crosses the ecliptic from south to north) and the angular distance between the ascending node and the perihelion. The problems dealt with here include passage from heliocentric to geocentric position or *vice versa*, with or without account being taken of aberration, nutation, and parallax.

Section 3 concerns 'relations between several places in orbit'. The most important problem resolved here is that of determining the orbit when two *radii vectores* are given in magnitude and position, with the time for the planet to describe the intermediate space. Gauss in *TM* develops entirely new formulas for this problem (arts. 87–105), eliminating the approximate integrals he had used earlier. In outline his procedure is as follows.

Let

$$f = \frac{v'}{2} - \frac{v}{2} \quad \text{and} \quad g = \frac{E'}{2} - \frac{E}{2}, \tag{26}$$

where v, v' are the true anomalies and E, E' the eccentric anomalies in the two positions. The angle f is known, being half the angle between the two *radii vectores*, whereas g is

unknown. Next let

$$\frac{\sqrt{\frac{r'}{r}}+\sqrt{\frac{r}{r'}}}{2\cos f}=1+2\ell; \tag{27}$$

here ℓ can be computed from the known quantities r, r', and f. Gauss shows that the semi-transverse axis will be

$$a=\frac{2(\ell+\sin^2\frac{1}{2}g)\cos f\sqrt{rr'}}{\sin^2 g}. \tag{28}$$

The difference in mean anomaly, $M'-M$, can be shown to be

$$\frac{kt}{a^{3/2}}=E'-e\sin E'-E+e\sin E=2g-\sin 2g+2\cos f\sin g\frac{\sqrt{rr'}}{a}, \tag{29}$$

where t is the known time for the planet to move from the first to the second position. Substituting the value of a from (28) in (29), and putting

$$\frac{kt}{2^{3/2}\cos^{3/2} f(rr')^{3/4}}=m, \tag{30}$$

which is evidently a known quantity, Gauss obtains

$$\pm m=\left(\ell+\sin^2\frac{1}{2}g\right)^{1/2}+\left(\ell+\sin^2\frac{1}{2}\right)^{3/2}\left(\frac{2g-\sin 2g}{\sin^3 g}\right). \tag{31}$$

Here the plus or minus sign is used according as $\sin g$ is positive or negative. In (31) g is the only unknown quantity. Note that m and the two terms on the right are proportional, respectively, to the area of the elliptical sector contained between the two *radii vectores*, the area of the triangle formed by the *radii vectores* and the chord, and the area of the elliptical segment cut off by the chord. The artifices whereby Gauss obtains g from (31) and then deduces the orbital elements from g are a high point of *TM* for sophistication and elegance.

In Section 4, concerning 'relations between several places in space', Gauss shows how to determine the inclination of the orbital plane to the ecliptic or other reference plane, and the longitude of its ascending node, from two complete observations, and develops systematically the consequences of the equations

$$0=nx-n'x'+n''x'', \qquad 0=ny-n'y'+n''y'', \qquad 0=nz-n'z'+n''z'', \tag{32}$$

these being the point of departure for all orbit-determinations.

6 SECTIONS 3 AND 4 OF BOOK II: THE METHOD OF LEAST SQUARES; PERTURBATIONS

In Section 3 Gauss undertakes to provide a derivation of the method of least squares. A.M. Legendre had described the method in his *Nouvelles méthodes pour la détermina-*

tion des orbites des comètes (1806), without offering a derivation. Gauss's derivation will gain wide acceptance in the 19th century, but [Gauss, 1823] will himself repudiate it.

Consider a theory depending on ν constants p_i to be determined observationally, and let μ observations M_j be applied to this purpose, where $\mu > \nu$. Were the p_i known, we could compute the values V_j to which the M_j should approximate. Let $V_j - M_j = v_j$. Gauss assumes a probability function $\varphi(v_j)$ exists giving the probability for v_j to have any particular value. Then (assuming the observations independent) the probability that the M_j have particular values v_j is

$$\prod_j \varphi(v_j) = \prod_j \varphi(V_j - M_j). \tag{33}$$

The most probable distribution of errors among the M_j can be obtained by maximizing (33), or equivalently, the function

$$\sum_j \log \varphi(v_j) = \Omega. \tag{34}$$

By Bayes's rule of inverse probability (§15.2), Gauss shows that, once the M_j are ascertained by observation, the maximization of (34) can be used to determine 'the most probable' values of the p_j; for each p_i we put

$$\partial\Omega/\partial p_i = 0, \quad \text{or} \quad \sum_j \frac{\partial v_j}{\partial p_i} \frac{1}{\varphi(v_j)} \frac{d\varphi(v_j)}{dv_j} = 0. \tag{35}$$

Here, however, we need an analytical formula for φ. Gauss makes the assumption that, if any quantity has been determined by several direct observations, made under the same circumstances and with equal care, the arithmetical mean of the observed values affords the 'most probable' value, and so obtains the formula

$$\varphi(v) = \frac{h}{\sqrt{\pi}} e^{-h^2 v^2}. \tag{36}$$

But in a later paper [Gauss, 1823] rejects this formula as merely hypothetical, and gives a new justification of the method based on the principle of minimal variance.

Section 4 deals with the role of perturbations in orbit determination. Gauss opposes taking them into account in an initial orbit-determination. After a good many observations have accumulated, perturbational analysis can become important for the refinement of orbital elements. It was so for Pallas and Juno, which are strongly perturbed by Jupiter.

7 THE IMPACT OF *TM*

TM received highly laudatory reviews from B.A. von Lindenau in *MC* (August 1809) and F.W. Bessel in *Jenaische Allgemeine Literatur-Zeitung* (April 1810). In Paris J.B.J. Delambre failed to comprehend Gauss's solution (31) of the sector-triangle problem (*Connaissance des temps*, 1812).

Four minor planets were identified between 1801 and 1807; the fifth was discovered in 1845. The frequency of discovery then increased dramatically, especially after introduction of photographic methods. Orbit-determination has thus remained a lively subject, and Gaussian procedures are still at the forefront, although allegiance to the Laplacian method (improved in important respects) has also persisted [Marsden, 1985, 1995]. Note from the beginning of this article that three translations of *TM* appeared in the 1860s.

TM's more general influence was to give a new emphasis to computational efficiency, to tracking error through calculations, and to the delicate fitting of theory to data. This Gaussian thrust had its chief initial effect among German astronomers, who under the leadership of Bessel, with the aid of German instrument-makers, brought astrometry to new levels of precision.

BIBLIOGRAPHY

Brandt, L. 1978. 'Über das Bahnbestimmungsproblem bei Gauss und Laplace. Eine Gegenüberstellung ihrer Methoden', *Mitteilungen der Gauss-Gesellschaft, 15*, 39–48.

Brendel, M. 1929. 'Über die astronomische Arbeiten von Gauss', in Gauss *Werke*, vol. 11, pt. 2, no. 3 (254 pp.).

Dunnington, G.W. 1955. *Carl Friedrich Gauss: titan of science*, New York: Hafner. [Repr. Washington: Mathematical Association of America, 2004.]

Gauss, C.F. 1809. 'Summarische Übersicht der zur Bestimmung der Bahnen der beiden neuen Hauptplaneten angewandten Methoden', *MC, 20*, 197–224. [Repr. in *Werke*, vol. 6, 148–165.]

Gauss, C.F. 1823. 'Theoria combinationis observationum erroribus minimis obnoxiae', *Commentarii Societatis Regiae Scientiarum Göttingensis, 5* (1819–1822), 33–60. [Repr. in *Werke*, vol. 6, 1–53.]

Laplace, P.S. 1799. *Traité de mécanique céleste*, vol. 1, Paris: Duprat. [Repr. as *Oeuvres complètes*, vol. 1, Paris: Gauthiers–Villars, 1878 (repr. Hildesheim: Olms, 1966). See §18.]

Marsden, B.G. 1985. 'Initial orbit determination: the pragmatist's point of view', *Astronomical journal, 90*, 1541–1547.

Marsden, B.G. 1995. 'Eighteenth- and nineteenth-century developments in the theory and practice of orbit determination', in *Planetary astronomy from the Renaissance to the rise of astrophysics.* Part B: *the eighteenth and nineteenth centuries*, Cambridge: Cambridge University Press, 181–190.

Reich, K. 1998. 'Gauss' Theoria motus: Entstehung, Quellen, Rezeption', *Mitteilungen der Gauss-Gesellschaft, 35*, 3–15.

Reich, K. 2001. *Im Umfeld der 'Theoria motus': Gauss' Briefwechsel mit Perthes, Laplace, Delambre und Legendre*, Göttingen: Vandenhoeck und Ruprecht.

CHAPTER 24

P.S. LAPLACE, *THÉORIE ANALYTIQUE DES PROBABILITÉS*, FIRST EDITION (1812); *ESSAI PHILOSOPHIQUE SUR LES PROBABILITÉS*, FIRST EDITION (1814)

Stephen M. Stigler

In the *Théorie* Laplace gave a new level of mathematical foundation and development both to probability theory and to mathematical statistics. The *Essai* brought the news to a relatively wide public.

Théorie analytique des probabilités

First publication. Paris: Courcier, 1812. 465 pages. Print-run: 1200 copies.

Later editions. 2nd 1814, 3rd 1820, both by Courcier. [3rd ed. repr. as *Oeuvres*, vol. 7, Paris: Imprimerie Royale, 1847; also as *Oeuvres complètes*, vol. 7, Paris: Gauthier–Villars, 1886 (photorepr. Hildesheim: Olms, 1966).]

Photoreprint of 1st ed. Brussels: Culture et Civilisation, 1967.

Essai philosophique sur les probabilités

First publication. As the Introduction to the 2nd edition of the *Théorie*.

First publication as a separate edition. Paris: Courcier, 1814. 97 pages. Print-run: 500 copies.

Photoreprint of 1st ed. Brussels: Culture et Civilisation, 1967.

Earlier versions. 1) As 'Séance 57 (21 Floréal)', *Séances des Écoles Normales, 6* (1795–1796) [repr. Paris: L'Imprimerie du Cercle-Social, 1800]. 2) Revised and enlarged as 'Notice sur les probabilities', *Annuaire du Bureau des Longitudes,* (1811: publ. 1810) [photorepr. in [Gillispie, 1979]]. 3) Further revised and enlarged as 'Séance 10' of 'Leçons de mathématiques données a l'École Normale, en 1795', *Journal de l'École Polytechnique, 2, cahiers 7–8* (1812).

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

Later editions. 2nd 1814, 3rd 1816, 4th 1819 (500 copies), all by Courcier. 5th 1825, 6th 1840, both Paris: Bachelier. 7th, Brussels: Société Belge de Librairie, 1840. 5th ed. also Brussels: H. Remy, 1829; also in *Oeuvres*, vol. 7, Paris: Imprimerie Royale, 1847, v–clxix; also in *Oeuvres complètes*, vol. 7, Paris: Gauthier–Villars, 1886, v–cliii (photorepr. Hildesheim: Olms, 1966); also (ed. and introd. X. Torau Bayle), Paris: É. Chiron, 1920; also 2 vols., Paris: Gauthier–Villars, 1921. Scholarly ed. based on the 5th ed. (ed. Bernard Bru), Paris: Christian Bourgois, 1986.

English translations. 1) *A philosophical essay on probabilities* (trans. F.W. Truscott and F.L. Emory, based on the 6th ed.), New York: John Wiley, 1902. [2nd ed. 1917, repr. New York: Dover, 1951.] 2) *Philosophical essay on probabilities* (trans. with notes by Andrew I. Dale, based on the 5th ed., indicating changes since the 1st ed.), New York: Springer-Verlag, 1995.

German translations. 1) *Philosophischer Versuch über Wahrscheinlichkeiten* (trans. F.W. Tönnies from the 3rd ed.), Heidelberg: Groos, 1819. 2) *Philosophischer Versuch über die Wahrscheinlichkeiten* (trans. N. Schwaiger from the 6th ed.), Leipzig: Duncker & Humblot, 1886. 3) 2) revised as *Philosophischer Versuch über die Wahrscheinlichkeit* (trans. H. Löwy with notes by R. von Mises), Leipzig: Akademische Verlagsgesellschaft, 1932 (*Ostwald's Klassiker der exakten Wissenschaften*, no. 233).

Spanish translation. Ensayo Filosófico sobre las Probabilidades (trans. and notes by Alfredo B. Besio and José Banfi), Buenos Aires: Espasa-Calpe, 1947.

Italian translation. Saggio filosofico sulle probabilità (trans. S. Oliva), Bari: Laterza, 1951.

Russian translation. Opyt filosofii teorii vieroiatnostei (ed. A.K. Vlasov), Moscow: Tipolit, 1908.

Related articles: Jakob Bernoulli (§6), De Moivre (§7), Laplace on celestial mechanics (§18), Pearson (§56), Fisher (§67).

1 THE MONT BLANC OF MATHEMATICAL ANALYSIS, AND ITS FOOTHILLS

In an 1837 review of the 3rd edition (1820) of Laplace's *Théorie analytique des probabilités* (hereafter, '*Théorie analytique*'), the British mathematician Augustus de Morgan wrote that 'Of all the masterpieces of analysis, this is perhaps the least known; [. . . it] is the Mont Blanc of mathematical analysis', he added, 'but the mountain has this advantage over the book, that there are guides always ready near the former, whereas the student has been left to his own method of encountering the latter' [de Morgan, 1837a, 347]. We could develop this metaphor further: the *Théorie analytique* emerged from a long series of slow processes and once established, loomed over the landscape for a century or more.

Three great treatises on probability had appeared towards the beginning of the 18th century; Pierre Remond de Montmort's *Essay d'analyse sur les jeux de hazard* 1708, 2nd ed. 1713, Jacob Bernoulli's posthumous *Ars conjectandi* (1713: see §6), and Abraham De Moivre's *Doctrine of chances* (1718, 2nd ed. 1738, 3rd ed. 1756; see §7). There were later short treatments, largely based upon De Moivre, by Thomas Simpson (1740, reprinted 1792), Samuel Clark (1758), and Charles F. Bicquilley (1783). The Marquis de Condorcet

published his monumental *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix* in 1785, a work that was influential in the history of mathematical social science but unimportant to the development of probability. The youthful A.-M. Ampère published a separate tract on the gambler's ruin problem in 1802. Other than these, and a handful of works of principal interest to gamblers, there was no new serious monograph on probability in any language between De Moivre's of 1718 and Laplace's of 1812. The intellectual landscape was not entirely barren of probability: there were a number of innovative memoirs by Simpson, Bayes, Lagrange, Daniel Bernoulli, and others. But in probability these were only foothills.

2 LAPLACIAN PROBABILITY

The biography of Pierre Simon Laplace (1749–1827) was reviewed briefly in §18.1. He published the first edition of *Théorie analytique* in 1812, at the age of 63 years. It represented the culmination of a professional lifetime of concern for the topic, and all of its text consisted of reworked versions of his earlier work. Laplace's prodigious abilities in the mathematical sciences were recognized early on, by his teachers in Normandy and by Jean d'Alembert in Paris when he was only 20. He was elected an Associate Member of the *Académie des Sciences* at age 24, already publishing deep and innovative memoirs on the mathematics of difference and differential equations, on the theory of gravitation and celestial mechanics, and on the theory of probability [Stigler, 1978]. He had contemplated writing a book on probability theory as early as 1780, when he published his second long memoir on the subject; but during the 1780s, he was (with a few notable distractions) drawn to problems in celestial mechanics. Inspired by work of J.L. Lagrange demonstrating the theoretical stability of the solar system, Laplace attacked the outstanding example of the three-body problem, showing that the observed inequalities in the motions of Jupiter and Saturn were consistent with the stability of the system. His success led him to undertake his magisterial *Traité de mécanique céleste* (1798–1805, 1823–1827: see §18).

Laplace's first important contribution to probability theory was a memoir [Laplace, 1774] on the probability of the causes of events. We would now describe this as Bayesian inference, although all evidence suggests it was entirely independent of Bayes's posthumously published 1763 essay, which seems to have gone unnoticed until the late 1770s (§15). In any case Laplace went far beyond Bayes and along quite different lines [Stigler, 1986a, 1986b]. He succinctly formulated the general problem: how could one learn about cause from effect; and how could the relative probabilities of an exhaustive list of causes be found from observing the effect and knowing the probability of that effect under each and every possible cause? Taking a uniform prior distribution over the possible causes for granted (as a 'Principe'), Laplace [1774] gave a general solution and examined in detail a binomial sampling model, including giving a demonstration of the asymptotic normality of the posterior distribution of the probability of success, a proof and result that foreshadowed much of his later work, including what has been called Laplace's method for the asymptotic approximation of definite integrals.

Laplace then addressed the problem of the mean; that is, when different astronomical observations of the same quantity are subject to error, how can they be best combined

to give a single determination of the quantity? He derived a symmetric exponential density from simple axioms and characterized the optimal estimate as that which minimized the posterior expected error, proving this was equivalent to choosing the median of the posterior distribution of the quantity. This would be a serious candidate as the first well-established result in mathematical statistics. He wrote two subsequent memoirs on these topics, one in 1777 that remained unpublished until resurrected in [Gillispie, 1979], and one in 1780 (published 1781).

Several of Laplace's memoirs of the 1780s on analysis (such as [1782] and [1785] on series and generating functions) were close to probability. Some of these were more explicitly probabilistic (on the asymptotic approximation of definite integrals); and in [Laplace, 1786] he published a short work on demography, where he developed the theory of the ratio method of estimating a population based upon records of births, deaths, or marriages, using a census of a few sample districts. But in the years after 1805 he returned to probability more systematically. In 1733 De Moivre had derived the normal probability density as an approximation to the binomial distribution (§7.6); in 1776 Lagrange had both generalized this to the multinomial distribution and shown how De Moivre's method of generating functions could be made more systematic and applied to sums of continuous random quantities in addressing the problem of the mean. In effect Lagrange had introduced the 'Laplace' transform, and given a small dictionary from which one might recognize the probability density of an arithmetic mean from its Laplace transform. In a few simple cases in 1785, and more definitively in [Laplace, 1810, 1811] he took this idea further, introducing 'Fourier' transforms where he showed how one could exploit the trigonometric representation of the complex exponential to arrive at an inversion formula for transforms.

In 1807–1809 Laplace worked on a long memoir that he must have seen as bringing his earlier work on asymptotic approximation of integrals and integral transforms to a logical conclusion, proving a quite general 'Central limit theorem' (to use the modern name, due to Georg Polya): a sum of a large number of independent random variables will have approximately a normal distribution, almost regardless of the distribution of the individual summands. It was a grand generalization of De Moivre's result, which was itself a central limit theorem when the summands were permitted to take only the values of either 0 or 1. The proof in [Laplace, 1811] was analytically a triumph of great power, even though, as de Morgan would later write of Laplace's mathematics in general, 'it gave neither finish nor beauty to the results'.

While that memoir was in press, already typeset for the *Académie*'s volume for 1809, Laplace was evidently startled to receive a copy of the treatise on celestial mechanics, *Theoria motus corporum coelestium* (1809) by C.F. Gauss (1777–1855) (§22). There he saw Gauss's development of the method of least squares and its connection to the normal distribution: Gauss showed that if for a simple sample the arithmetic mean was optimal, then the errors must be normally distributed, and in that case the general optimality of least squares must follow. The proof was elegant but the assumption by hypothesis of the superiority of the arithmetic mean would have been uncongenial to Laplace (he had proved that for other distributions the arithmetic mean was *not* best as early as 1774). Still, this derivation of the method of least squares, with the central point being the relationship of the method to the normal distribution, must have hit Laplace like a bolt. He rushed a short supplement to his paper to the press; it appeared at the very end of the *Académie*'s volume

for 1809, published in 1810. In the supplement [Laplace, 1810] restated his central limit theorem more elegantly than in the original memoir, cited Gauss, and showed how his theorem provided a much more palatable basis for the assumption of normally distributed errors: each error might itself be a sum of more elementary constituents. In later literature this would be called the hypothesis of elementary errors. With his supplement safely to press, Laplace went to work more seriously. In 1810–1811 he produced a further memoir [Laplace, 1811] that remade error theory and would form the core of the statistical portion of *Théorie analytique*.

3 PUBLICATION OF THE *THÉORIE ANALYTIQUE*

The contents of the first edition (1812) of the *Théorie analytique* are summarised in Table 1. The book carried an effusive dedication, 'A Napoléon-le-grand'. Laplace probably first met Napoleon Bonaparte in the 1780s, when he would have examined him at the *Ecole Militaire*; in 1798 Napoleon had appointed Laplace to be Minister of the Interior, immediately after the Coup of the 18 *Brumaire* (when Napoleon ousted the ruling Directory, replacing it with the *Consuls de la Republique* with himself as chief Consul). But that appointment only lasted six weeks, after which Laplace was replaced by Napoleon's brother Lucien. Evidently, once power was consolidated there was no need for a prestigious but inexperienced scientist in the government. In his memoirs written at Saint Helena, Napoleon tendentiously stated that Laplace had brought the spirit of the infinitesimal into governmental councils, but there seems to be no doubt that he was only appointed as a short-term figurehead, a place-holder while Napoleon consolidated power.

The book *Théorie analytique* itself was a mixture of old and new. Book One was a lightly rewritten version of earlier memoirs; Part I was little altered from a 1782 memoir on series, while Part II was derived from memoirs of 1785, 1786, and other early publications. Intellectually the organization could be defended: the mathematical tools were developed before they were deployed. But strategically the arrangement was unfortunate. Any reader who started with Book One without a talent and a taste for extremely difficult mathematics would likely have put the book aside before encountering any discussion of probability. On the other hand, a reader who followed a guide such as de Morgan would recommend and skipped to Book Two, prepared to go lightly over some sections at first reading—that reader would be rewarded with an astonishing advance over anything that had come before.

4 LAPLACE'S *THÉORIE* AND MATHEMATICAL STATISTICS

Readers of De Moivre's *Doctrine of chances*, whose posthumous third edition had appeared over half a century before, would have recognized some parts of Laplace's Chapters 1 and 2 of Book Two, but little else in the entire work. Some past results of Laplace's own were simply dusted off and presented cryptically, such as his derivation from 1780 of the expectation of a cleverly defined random probability density,

$$E\big[f(x)\big] = \frac{1}{2a}\log\left(\frac{a}{|x|}\right), \quad \text{for } |x| \leqslant a, \tag{1}$$

Table 1. Summary by Chapters of *Théorie analytique des probabilités* (1812). Most titles are translated.

Ch. (pp.)	**Topics**
Book I	*Calculus of generating functions.*
Part I	*General considerations.*
1 (46)	Generating functions of one variable: interpolation of series, integration of linear differential equations, transformation of series.
2 (38)	Generating functions of two variables: interpolation of series in two variables, integration of linear partial difference equations, equations in many variables, passage from finite differences to infinitely small differences.
Part II	*Theory of asymptotic approximations.*
1 (21)	Approximation of integrals of factors raised to high powers.
2 (26)	Approximate integration of linear equations in finite and infinitely small differences.
3 (51)	Various applications of the preceding methods.
Book II	*General theory of probabilities.*
1 (12)	General principles of this theory: definitions and assumptions.
2 (86)	The probability of events composed of simple events whose possibilities are given: lotteries, balls and urns, games, a derivation of a symmetric error distribution.
3 (29)	Laws of large numbers: central limit theorems.
4 (45)	The probability of errors in a mean taken of a large number of observations, and the most advantageous method of taking a mean. Inversion of Fourier transforms, the asymptotic normality of linear estimators, the large sample optimality of least squares, the bivariate asymptotic normality of pairs of linear estimators and the optimality of least squares in that case, asymmetric error distributions, the problem of choosing a mean investigated from the a posteriori point of view.
5 (14)	The application of the calculus of probabilities: to the probability of subtle differences in meteorology or astronomy or physiology, to Buffon's needle.
6 (39)	The probability of causes or of future events; ratio estimation of population size.
7 (6)	The effect of unknown deviation from equality of probabilities of composite events.
8 (11)	The mean duration of lives, of marriages, and of associations of more than two people.
9 (13)	Some consequences of the probability of future events for computing mortality tables and insurance calculations.

Table 1. (*Continued*)

Ch. (pp.)	**Topics**
10 (14)	On moral expectation; logarithmic utility and Daniel Bernoulli's application to the St. Petersburg problem.
11 (added in 2nd ed.) (16)	The application of probability to questions of evidence and testimony.
Additions I–III in 2nd ed. (23)	Wallis's infinite product for π, and two direct demonstrations of results from Book I.

that he suggested as a candidate for an error distribution at the end of Chapter 2. In Chapter 3 he reprised some of the analysis of his earlier memoirs in 1785 and 1786 on the normal approximation to the integrals of functions raised to high powers, with applications. But with Chapter 4 there was no more rehashing of ancient work; the mature Laplace was now in full stride.

In the previous year Laplace had published an extensive memoir on what we would now call the asymptotic theory of linear estimation, building on the ideas that had been inspired by his reading of Gauss. Chapter 4 gave a vigorous re-presentation of this work, and it was a tour de force. Taking advantage of the development of Book One, he was able to present his central limit theorem in just a few pages, for the case of a sum of independent identically distributed errors, for a general discrete error distribution. That part of the text could be read as a tutorial for the use of the Fourier transform in probability, preceding first for a symmetric uniform distribution over a bounded set of integers, where trigonometric identities gave a simple solution via his inversion formula, to the more general case, where another step of asymptotic approximation was required. And with the case of identically distributed summands established, he moved on to deal with general weighted sums of the form $\sum_i a_i Y_i$.

We can succinctly summarize Laplace's procedure as follows, in modernized notation. Let $Y_i = \alpha X_i + \varepsilon_i$, where the X's are taken as known, fixed numbers, the Y's as observed, α as unknown and to be estimated, and the ε_i as random errors in the observations, supposed symmetrically distributed about zero. Then any linear estimator of α of the form

$$\frac{\sum m_i Y_i}{\sum m_i X_i} \quad \text{will estimate } \alpha \text{ with error} \quad \frac{\sum m_i \varepsilon_i}{\sum m_i X_i}, \tag{2}$$

itself symmetrically distributed about zero. Since such weighted sums were shown to be asymptotically normally distributed, with mean error zero, one could compare choices of weights m_i simply by comparing the variances of the limiting distributions. As Laplace noted, the scheme with the smallest variance would have the smallest probability of being outside any error bounds you might choose. And, as he showed, the weights given by the method of least squares, namely $m_i = X_i$ would give that smallest variance! He went on to show that the same property of least squares would hold for linear models with two unknowns, that is,

$$Y_i = \alpha X_i + \beta W_i + \varepsilon_i. \tag{3}$$

This included finding the bivariate asymptotic normal distribution of the two estimators. He then showed that the same optimality would hold for asymmetric error distributions. All of this was accomplished in only 28 pages.

It is worth pausing to emphasize the novelty in this work. Gauss had noted in 1809 that when error distributions were exactly normal, least squares estimators would maximize what we now call the likelihood function—the density of the observed Y's. Nearly a decade later in two memoirs of 1821 and 1823, Gauss would cast off his earlier approach and prove what is now called 'the Gauss–Markov theorem': that among all linear estimators whose mean error was zero, the least squares estimators had smallest variance regardless of the error distribution. Laplace had shown a result that was nearly the same as this latter one and in some respects more satisfactory: among all linear estimators whose mean error was zero, the least squares estimators had smallest variance regardless of the error distribution *if the number of observations was quite large*. The conclusion was weaker in being only asymptotic; it was stronger in that because Laplace could conclude the least squares estimators were normally distributed, they would be superior to any other linear estimator in the class in *every* sense, not merely as having smaller mean squared error. In that respect he improved upon Gauss's later memoirs a decade prior to Gauss's work!

In the remaining 15 pages of this remarkable chapter, Laplace reconsidered the linear estimation problem from a Bayesian perspective. Echoing his memoir of 1774 but in a much more complex setting, he showed the posterior distribution of the coefficients was asymptotically normal, and that for this asymptotic distribution least squares was again optimum.

After Chapter 4, the remainder of *Théorie analytique* seems anticlimactic. There were strong results for special problems, but the methods were no longer a surprise to the reader, and many of the results reprised those of memoirs from 1774 to 1786. A special problem in astronomy was treated in Chapter 5, and a generalization of Buffon's needle was presented. Laplace's analysis of the asymptotic theory of the ratio estimator of population from a Bayesian perspective can be found in Chapter 6. In his 1774 memoir he had noticed that uncertainty about the balance of a coin (what if the probability of a face was not known to be $1/2$, but only known to be in the interval $[1/2-\varepsilon, 1/2+\varepsilon]$?), could have an unforeseen effect upon the probabilities of complicated bets. This was traced out here in Chapter 7. Chapter 8 analyzed mortality tables; Chapter 9 looked at life insurance. And in Chapter 10, the last of the first edition, he explored the use of Daniel Bernoulli's logarithmic utility functions in problems like the St. Petersburg Paradox.

In 1814 Laplace issued a second edition with three changes: the dedication to Napoleon was removed (reflecting the changed political climate), a 106-page introduction was included, and a chapter on the application of probability to questions of evidence and testimony and three short technical appendices were added. Two supplements, on the application of probability to the natural sciences and to geodetic operations, were issued in 1816 and 1818. In 1820 the third and final edition was published; it differed from the second by the addition of the two previously issued supplements and a third supplement on the application of probability to surveying. In 1825 a fourth supplement appeared, written mostly by Laplace's son; it was included with copies of the third edition sold after 1825.

5 THE *ESSAI PHILOSOPHIQUE*

The long introduction that Laplace added to the second edition was also published separately in the same year, as *Essai philosophique sur les probabilités*. Its contents are summarised in Table 2. This *Essai* was extremely popular and influential; unlike the *Théorie analytique,* it required no guide and was widely read and quoted. The *Essai* was itself the product of years of thought by Laplace; four different versions saw publication even before the first edition appeared! It was to the *Théorie analytique* much like his 1796 *Exposition du systême du monde* was to the *Mécanique céleste* (§18): it presented the ideas of the *Théorie analytique* to a broad popular audience while serving as a synopsis of the full treatise, in part mirroring it in organization. Much of the *Essai* is as elegantly written as any work in popular mathematics, although translations seldom reflect that elegance. It was in the *Essai* that Laplace gave his famous statement of what has been called Laplacian determinism, that physical laws in principle predetermined all physical action at all levels for all time (p. 2):

> If an intelligence, at a given instant, knew all the forces that animate nature and the position of each constituent being; if, moreover, this intelligence were sufficiently great to submit these data to analysis, it could embrace in the same

Table 2. Summary by Parts of *Essai philosophique sur les probabilités* (1814). The titles are translated.

Pages	**Topics**
1–36	*Philosophical essay on probabilities*.
2–7	On probability.
7–19	General principles of the calculus of probabilities.
19–22	On expectation.
22–36	Analytical methods of the calculus of probabilities.
37–96	*Applications of the calculus of probabilities*.
37–38	On games.
38–40	Unknown inequalities that can exist among probabilities supposed equal.
41–49	The laws of probability which result from the indefinite repetition of events.
49–60	The application to research on phenomena and their causes.
60–64	Choosing a mean among observations.
64–70	Tables of mortality, and the mean durations of life, marriages, and other associations.
70–73	Gains and losses depending upon the probability of events.
73–76	The choices and decisions of assemblies.
77–83	Some illusions in the estimation of probabilities.
83–89	The different means of approaching certainty.
89–96	The history of probability.

> formula the movements of the greatest bodies of the universe and those of the smallest atoms: to this intelligence nothing would be uncertain, and the future, as the past, would be present to its eyes.

And two pages later he reconciled this with the application of probability to physical phenomena:

> The regularity which astronomy shows us in the movements of the comets doubtless occurs in all phenomena. The curve described by a simple molecule of air or water vapor is regulated in a manner just as certain as the orbits of the planets; the only difference between these is that introduced by our ignorance. Probability is relative in part to this ignorance, and in part to our knowledge.

In some places Laplace was defeated in his attempt to convey a sense of mathematical objects in prose. His verbal description of a generating function was opaque to a non-mathematician but at least descriptive. But what would a non-mathematician make of his evocation of the normal density? After struggling with this problem he finally settled on the following in the third edition: 'The probability of the errors remaining in each element is proportional to the number whose hyperbolic logarithm is unity, raised to a power equal to the square of the error, taken as negative, and multiplied by a constant coefficient which can be considered as the modulus of the probability of the errors' (p. 92).

Five separate editions of the *Essai* were published in France during Laplace's life, two in 1814 (the first in quarto, the second octavo), a third in 1816, a fourth in 1819 (serving also as the introduction to the third edition of *Théorie analytique*), and a fifth in 1825. An unaltered posthumous sixth edition appeared in 1840. Each of the first five editions shows extensive revision as well as the addition of new material; for example, the third edition added a discussion of the relationship of associationist psychology to errors of observation. A thoroughly annotated scholarly edition based upon the fifth edition of 1825 was published in 1986, edited by Bernard Bru.

Laplace's summary of the *Essai*, and indeed of his body of work on probability, may not be accurate as judged by those few who have actually read it all; but it has struck a resonating chord in generations of students and philosophers since (p. 95):

> We have seen in this essay that the theory of probabilities is essentially only common sense reduced to calculation; it helps us to judge accurately what sound minds perceive by a sort of instinct, often without being able to give a reason.

6 THE LEGACY

The *Théorie analytique* was widely known and immensely influential, but it was not widely read, and it probably sold poorly. The second and even the third editions were evidently cobbled together from unsold copies of the first edition. The dedication to Napoleon of course had to be dropped in 1814, and that may have been the impetus for what few additional changes were made. Starting with the second edition there were new title pages and the long introduction, a single added chapter and appendix, and, as noted

in [Todhunter, 1865, 495–497], seven cancelled replaced pages, those containing embarrassing misprints (such as the over-estimate of the birthrate on p. 391, leading to a consequent 50% excess for the population of France: 42,529,267 instead of the more reasonable 28,352,845). The third edition added three supplements and revised the introduction, but the text was otherwise unaltered, and bound copies, even those including the 1825 fourth Supplement, still betray the cosmetic nature of the only changes in the form of the stubs of pages inserted for the seven cancels. This suggests that the first printing was too large, or the sales quite small, or both of these. Has any other great work gone through three editions with so little change?

Rather than wide sales (or even, given its difficulty, wide readership), the book became known through its effect upon a small number of major scholars, some of whom produced more accessible treatments. The first of these was S.F. Lacroix's *Traité élémentaire du calcul des probabilités*, published in 1816, with several later editions. Other available French treatises were written by Siméon Poisson in 1837 and Augustin Cournot in 1843 (himself influenced by Lacroix, as Bru has documented); and Irénée Jules Bienaymé developed Laplace's theory further in a number of important memoirs from 1838–1852 [Heyde and Seneta, 1977]. In England, Laplace's *Théorie analytique* was a standard reference at Cambridge, and de Morgan based his *Encyclopedia Metropolitana* article [1837b] on the *Théorie analytique*; the separate edition of that article in 1849 could be considered a treatise. Victor Bunyakovsky's book of 1846 made Laplace's work available in Russia, where it influenced Michel Ostrogradsky and Pafnuty Chebyshev.

Gauss had been a catalyst to the surge of Laplace's activity that had produced the *Théorie analytique*, and Gauss himself would have been one of its more important readers. Indeed, his reconsideration of least squares between 1821 and 1823 can be best interpreted as a reaction to the new developments that Laplace assembled in the *Théorie analytique*. The two approaches, Laplace's with his emphasis upon asymptotic approximation and Gauss's with his preference for exact mathematics, were mutually reinforcing at the time and dominated the discussions of mathematical statistics for a century. In probability, Laplace's *Théorie analytique* and *Essai* stood as unchallenged beacons well into the 20th century, even if the influence of the former was generally through secondary accounts and subsequent extensions by others.

BIBLIOGRAPHY

de Morgan, A. 1837a. Review of *Théorie analytique des probabilités*, *The Dublin review, 2*, 338–354; *3*, 237–248.

de Morgan, A. 1837b. *Theory of probabilities*, London: John Joseph Griffin. [Article prepared for the *Encyclopædia Metropolitana* and reissued separately in 1849.]

Gillispie, C.C. 1979. 'Mémoires inédits ou anonymes de Laplace sur la théorie des erreurs, les polynômes de Legendre, et la philosophie des probabilités', *Revue d'histoire des sciences, 32*, 223–279. [Includes the first printing of Laplace's 1777 'Recherches sur le milieu qu'il faut choisir entre les résultats de plusieurs observations'.]

Gillispie, C.C. 1997. *Pierre-Simon Laplace, 1749–1827*: *a life in exact science*, Princeton: Princeton University Press.

Heyde, C.C. and Seneta, E. 1977. *I.J. Bienaymé: statistical theory anticipated*, New York: Springer-Verlag.

Laplace, P.S. *Works. Oeuvres complètes*, 14 vols., Paris: Gauthier–Villars, 1878–1912. [Repr. Hildesheim: Olms, 1966.]

Laplace, P.S. 1774. 'Mémoire sur la probabilité des causes par les évènemens', *Mémoires de mathématique et de physique, presentés à l'Académie Royale des Sciences, par divers savans, & lûs dans ses assemblées, 6*, 621–656. [Repr. in *Works*, vol. 8, 27–65. Translated in [Stigler, 1986b].]

Laplace, P.S. 1782. 'Mémoire sur les suites', *Mémoires de l'Académie Royale des Sciences de Paris* (1779), 207–309. [Repr. in *Works*, vol. 10, 1–89.]

Laplace, P.S. 1785, 1786. 'Mémoire sur les approximations des formules qui sont fonctions de très-grands nombres' and 'Suite', *Ibidem*, (1782), 209–291; (1783), 423–467. [Repr. in *Works*, vol. 10, 209–291; 295–338.]

Laplace, P.S. 1810. 'Mémoire sur les approximations des formules qui sont fonctions de très-grands nombres, et sur leur application aux probabilités', *Mémoires de la classe des sciences mathématiques et physiques de l'Institut de France*, (1809), 353–415; 'Supplément', 559–565. [Repr. in *Works*, vol. 12, 301–353.]

Laplace, P.S. 1811. 'Mémoire sur les intégrales définies, et leur application aux probabilities, et spécialement à la recherché du milieu qu'il faut choisir entre les resultants des observations', *Ibidem* (1810), 279–347. [Repr. in *Works*, vol. 12, 357–412.]

Stigler, S.M. 1978. 'Laplace's early work: chronology and citations', *Isis*, *69*, 234–254.

Stigler, S.M. 1986a. *The history of statistics: the measurement of uncertainty before 1900*, Cambridge, MA: Harvard University Press.

Stigler, S.M. 1986b. 'Laplace's 1774 memoir on inverse probability', *Statistical science*, *1*, 359–378.

Todhunter, I. 1865. *A history of the mathematical theory of probability from the time of Pascal to that of Laplace*, London: Macmillan. [Repr. New York: Chelsea, 1949, 1965.]

CHAPTER 25

A.-L. CAUCHY, *COURS D'ANALYSE* (1821) AND *RÉSUMÉ* OF THE CALCULUS (1823)

I. Grattan-Guinness

In these two books Cauchy laid out a theory of limits, and upon its basis he constructed the basic theory of real-variable functions and of the convergence of infinite series; and also the calculus, in the approach that eventually was to dominate all others.

Cours d'analyse

First publication. Cours d'analyse de l'Ecole Polytechnique Royale 1re Partie. Analyse algébrique. Paris: de Bure, 1821. xvii + 576 pages. [No more parts published.]

Photoreprints. Darmstadt: Wissenschaftliche Buchgesellschaft, 1968. Paris: Gabay, 1989. Bologna: CLUEB, 1992 (introd. by U. Bottazzini).

Reprint. As *Oeuvres complètes*, series 2, volume 3, Paris: Gauthiers–Villars, 1897.

German translations. 1) *Lehrbuch der algebraischen Analysis* (trans. C.L.B. Huzler), Königsberg: Bornträger, 1828. 2) *Algebraische Analysis* (trans. C. Itzigohn), Berlin: J. Springer, 1885.

Russian translation. Algebraicheskie analiz (trans. F. Ewald, B. Grigoriev and A. Ilin), Leipzig: 1864.

Résumé

First publication. Résumé des leçons données à l'Ecole Polytechnique Royale sur le calcul infinitésimal. Tome premier. Paris: de Bure, 1823. 182 pages. [No more volumes published.]

Photoreprints. Paris: ACL Editions, 1987; Paris: Ellipses, 1994. [Lack the second addition.]

Reprint. As *Oeuvres complètes*, series 2, volume 4, Paris: Gauthiers–Villars, 1898, 5–261.

Part Russian translation. Kratkoe izlozhenie urokov ob differentsal'nom i integral'om ischislenii prepodavlaemikh v korolevskoi politekhnicheskoe shkole (trans. B.Ya. Bunyakovsky), Saint Petersburg: 1831.

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

Related articles: Lagrange on the calculus (§19), Lacroix (§20), Cauchy on complex-variable analysis (§28), Fourier (§26), Riemann on trigonometric series (§38), Cantor (§46).

1 FROM STUDENT TO PROFESSOR

Cauchy's father was Secretary of the Senate, and came to know J.L. Lagrange and P.S. Laplace as senators; so these senior mathematicians were aware early on of the talented son. He entered the *Ecole Polytechnique* in 1805, and after the two-year course proceeded in 1807 to the *Ecole des Ponts et Chaussées* for a further three years. During this time he had direct contact with Gaspard Riche de Prony (1755–1839), as professor at the first school and Director of the second. Following the usual practice, Cauchy then entered the *Corps des Ponts et Chaussées* as an engineer, and was involved with various projects, including the large canal and port system being constructed in Paris. But his mathematical researches had already started, and his career was to gain artificial boosts with the fall of Napoléon in 1815.

It is a great irony that Cauchy was born in 1789, the year of the French Revolution; for until his death in 1857 he held unswervingly to the Bourbon monarchy and to the Catholic faith that they upheld. Thus, upon their Restoration in 1815 he profited greatly from his adherence. A reform of the *Ecole Polytechnique* included his appointment as a professor there, along with fellow Catholic A.M. Ampère; they replaced his former professors de Prony and S.D. Poisson (1781–1840), who became the graduation examiners for mathematics. Further, the *Académie des Sciences* was restored under its pre-revolutionary name (after functioning since 1793 as a class of the *Institut de France*), and in an event unique in its history Cauchy and the clock-maker A.L. Bréguet were appointed *without election* to replace two dismissed colleagues of Napoléon, Gaspard Monge and Lazare Carnot. Cauchy was also made an adjunct professor at the *Faculté des Sciences* in Paris, and sometimes he substituted as Professor of Physics at the *Collège de France.*

Everything was going extraordinarily well for Cauchy during the Restoration. Indeed, in a manner showing few parallels the man and the mathematician were in total harmony: God in heaven, the King on the Throne, truths in science, and (especially for our context) rigour in mathematics as much as possible. He even introduced the systematic numbering of formulae in an article or a chapter of a book. During the Bourbon reign from 1816 to 1830 he produced material that takes up about 12 of the 27 quarto volumes of his collected works, and most of the best stuff. He became so prolific that from 1826 he published *Exercices de mathématiques*, his own journal in that he was sole author: averaging about 32 pages per month, it appeared regularly until 1830, for a reason explained in Section 5.

Among biographies the most valuable is [Belhoste, 1991]. The older one, [Valson, 1868], is compromised by Catholic hagiography; but it contains valuable information, including on manuscripts in his *Nachlass*, most of which was destroyed by the family in 1937 when the *Académie des Sciences* refused to accept it as a collection.

Cauchy's marriage in 1818 into a publishing family of de Bure ('Booksellers of the King') even provided the outlet for his books, including the two under discussion here. Both were based upon his teaching at the *Ecole Polytechnique*, which is reported in detail in [Gilain, 1989]. He and Ampère taught their respective cohorts for both years, Cauchy

starting in even-numbered years. On the books and their context see [Grattan-Guinness, 1990, chs. 10–11] and [Bottazzini, 1992]. The *Cours d'analyse* covered limits and continuity, functions and infinite series, while the *Résumé* continued with the calculus. The reaction of the school will be noted in section 6, when the impact and Cauchy's own career after 1830 are reviewed.

During this period Cauchy was also developing complex-variable analysis, which appears to a small extent in both books; an account is provided in (§28). Both subjects are discussed in [Bottazzini, 1986, chs. 2–5].

2 THE *COURS D'ANALYSE*: 'ALGEBRAIC ANALYSIS' AND THE THEORY OF LIMITS

Table 1 summarises the *Cours d'analyse*, using the original page numbers; to estimate those in the *Oeuvres* edition subtract about 20%. Early in his introduction Cauchy characterised 'algebraic analysis' as concerning types of (real- and complex-) variable functions, convergent and divergent series, 'the resolution of equations, and the decomposition of rational fractions' (pp. i–ii). Later he stated some of the main features; curiously he did not refer to the theory of limits, which grounded his approach. It appeared in the 'preliminaries', where some opening chat about number and quantity was followed by the notion of variable, and this crucial definition: 'When the values successively attributed to the same variable approach indefinitely a fixed value, so as to differ from it as little as one might wish, this latter is called the *limit* of all the others' (p. 4). Clearly he intended passage both over real numbers and over integers, and, more importantly, with no restriction over the *manner* of passage: by zig-zag around the limiting value as well as approach solely from below and solely from above. He adopted the symbol '*lim*.' (p. 13), which had been introduced in 1786 by the Swiss mathematician Simon l'Huilier, whose theory of limits was rendered both cumbersome and constrained by separate treatments of (only) monotonic passage.

Cauchy also noted that more than one limiting value might obtain, when a double bracket was used: thus 'arc.sin$((a))$' denoted all values of this function, with 'arc.sin (a)' for the smallest value (pp. 7–8). He then ran through cases of values for various simple functions; and in Note 1 he rehearsed the properties of real quantities at a length surprising for a book at this level. Twice there he mentioned of an irrational number that 'one can obtain it by rational numbers of ever more approximate values' (pp. 409, 415); but he did not envision this as a *definition.*

Cauchy's use of limits gave great status to *sequences* of quantities; and one of his main techniques in proof was to show the existence of mean values between the maximum and minimum of a sequence. He proved various existence theorems for given sequences (for example on p. 15, that $(\sum_r a_r / \sum_r b_r)$ lay within the range of values of the sequence $\{a_r/b_r\}$); Note 2 elaborated upon the feature. Such theorems helped to secure the basic properties of limits, such as preserving the arithmetical operations (lim sum = sum lim, and so on). He also surpassed his predecessors in recognising, and as a fundamental matter, that a limiting value might *not* obtain; for example and importantly, 'a *divergent series does not have a sum*' (p. iv).

Cauchy devoted ch. 2 to a topic which may surprise: 'infinitely small and infinitely large quantities'. They were defined as variables passing though sequences of values which took

Table 1. Contents of the *Cours d'analyse*. Chapters are followed by Notes.

Ch./ Note	Page	Topics
		'Introduction' (8 pages): aims, some main features.
		'Preliminaries': real quantities; mean values.
1	1	'On real functions': general; simple and compound.
2	19	'Infinitely small and large quantities'; continuous functions; singular values.
3	26	'On symmetric and alternating functions; first-order equations.
4	70	'Determination of entire functions' (polynomials); interpolation.
5	85	'Determination of continuous functions' in one variable.
6	103	Convergent and divergent real series; tests; some summations.
7	123	'On imaginary expressions and their moduli'.
8	173	'On derivatives and imaginary functions' (of, for example, variables).
9	240	'On imaginary convergent and divergent series'; some summations.
10	274	'On the real and imaginary roots of algebraic equations'.
11	329	'Decomposition of rational fractions'.
12	365	'On recurrent series'.
1	380	'On the theory of positive and negative quantities'.
2	403	Formulae involving inequalities and mean values of quantities.
3	438	'On the numerical resolution of equations'.
4	460	Expansion of an alternating function.
5	521	Lagrange's interpolation formula extended to rational functions.
6	525	'On figured numbers' (binomial coefficients).
7	530	'On double series': convergence.
8	537	Expansions of multiple (co)sines in power series of (co)sines.
9	548	Infinite products: definition, convergence. [End 576.]

respectively zero and infinity as limiting values: $1/4$, $1/3$, $1/6$, $1/5$, ... was (his only) example of the first case, and $1, 2, 3, \ldots$ of the second. Then he proved various theorems on the orders of infinitude to which variables may be subject. The mathematics as such is unobjectionable; but his use of the adverb 'infinitely' is unfortunate, especially for a student reader, since these quantities have no necessary connection with infinitesimals as used in the Leibnizian calculus (compare §4.2).

3 THE *COURS D'ANALYSE*: CONTINUOUS FUNCTIONS AND INFINITE SERIES

Cauchy's next topic was 'the continuity of functions': $f(x)$ '*will remain continuous with respect to the given limits*' x_0 and X $(> x_0)$ '*if, between these limits, an infinitely small increase of the variable always produces an infinitely small increase of the function itself*', with a re-statement for continuity '*in the vicinity of a particular value of the variable* x' (pp. 34–35). Various theorems followed easily, for functions of one and several variables. However, the intermediate value theorem, that

$$\text{for any value } b \text{ such that } f(x_0) < b < f(X),\ \ f(x) = b \tag{1}$$

for at least one value of x between x_0 and X, was problematic. Uncharacteristically, in the proof he described a 'curve' given by $y = f(x)$ changing values in the plane, and inevitably hitting the line $y = b$ at least once (pp. 43–44). A more refined proof was given in Note 3, where he partitioned the range of values of x and sought the zero by successive approximation; but, as with his remark above about irrational and rational numbers, the *existence* of the required limiting value was assumed rather than established.

Cauchy also proved various theorems on the comparative rates of increase or decrease of values of functions. They came into play in ch. 5 on the convergence of infinite series, specified in terms of the passage of the nth partial sum 's_n' to value s as n increased and the '*remainder*' term 'r_n' approached zero as its limiting value (pp. 123–131): I quote his terms and symbols because he popularised them here. One main consequence were convergence tests, sufficient conditions upon the terms for convergence (and divergence) to occur. He gave the tests which we now call root, ratio, condensation, logarithmic, product and alternating, and proved them for series with the terms of the same and of mixed signs. In a somewhat vague way his tests used upper limits of sequences of values rather than simple limiting values. Some results had been known before, but usually against an imperfect understanding of convergence itself and often concerned with (supposed) *rates* of convergence. The theory was complemented by comparable studies of double series and infinite products (Notes 7, 9). Elaboration of these tests, and the brother theorems on functions, constituted a notable part of the influence of this book [Grattan-Guinness, 1970, appendix].

But in this chapter also lay trouble, or at least a theorem which has led to much discussion of Cauchy's doctrine: if an infinite series of continuous functions converged to s in the vicinity of some value of x, then s was also continuous there. The proof worked by showing that when x increased by an 'infinitely small' amount, then the incremental increase or decrease on both s_n and r_n was small, and thus that it was also small on s (pp. 131–132). The trouble is that, for example, Fourier series (then new: see §26) seemed to contradict this theorem when the represented function was discontinuous. Some commentators, such as [Laugwitz, 1987], have defended Cauchy's proof on the grounds that he used sequences of values via his construal of infinitesimals, and also that he defined continuity over the interval of values of x between x_0 and X, and did not specify point-wise continuity. My view is that neither here nor elsewhere did he grasp the special needs of *multiple*-limit processes required by the distinction between point-wise continuity and continuity over an interval, and that this theorem involves it: x varies but so does n, and the relationship between the two processes needs careful attention which came only decades later with Karl Weierstrass and his followers [Grattan-Guinness, 1970, ch. 6]. In section 4 I shall mention another theorem where such refinements are lacking.

For the binomial theorem Cauchy broke with tradition in establishing it independently of the calculus; instead he used functional equations, which had gained status recently especially because of their use in Lagrange's conception of analysis. He found that the solution for continuous functions ϕ of

$$\phi(x + y) = \phi(x)\phi(y) \text{ was } \phi(x) = Ax,\ A \text{ constant} \tag{2}$$

(pp. 107–109). Then he showed that the binomial series $(1+x)^n$ (n not necessarily integral) also satisfied $(2)_1$ for $|x| < 1$ (pp. 165–172).

Some aspects of complex-variable analysis were also presented, as much as possible by analogy with real variables; for example, continuity of a function and convergence of series (chs. 8–9). Complex ('imaginary') numbers themselves, construed algebraically (ch. 7), played a central role in proof of the fundamental theorem of algebra (ch. 10). Effected on the polynomial equation $f(x) = 0$ of degree n, with x and all coefficients complex (maybe real), it fell into three parts: show that all the roots were real or complex; hence that it could be written as a product of linear complex (maybe some or all real) factors; and that the number of these factors equalled the degree of $f(x)$. The proof relied much on the continuity of the real and imaginary parts of $f(x)$ formed after converting x into complex polar coordinates; being positive in value under certain circumstances and negative in others, then, by the intermediate value theorem, they took the value zero for at least one value of x. He was influenced by the second (1816) proof by C.F. Gauss, but with continuity defined his own way.

4 THE *RÉSUMÉ*: A NEW VERSION OF THE CALCULUS

Apart from some minor anticipations by, for example, Ampère, Cauchy's treatment of quantities, functions and infinite series in the *Cours d'analyse* was revolutionary; that is, other views had to go. Among those, his special target was Lagrange's founding of the calculus on Taylor's series, and indeed the whole aim of algebraising mathematics (§19). In his introduction Cauchy wished to give methods 'all the rigour that one demand in geometry, so as never to draw upon reasoning drawn from the generality of algebra' (p. ii). He admired here the (supposed) rigour of Euclid, not geometry as such; *like* Lagrange, he also gave no diagrams in his book, for his approach was not grounded in either geometry or algebra. He was presenting 'mathematical analysis' (p. v), the umbrella discipline based upon limits and mean values and careful definitions and arguments based upon them.

This change from Lagrange is especially clear in the *Résumé* of Cauchy's lectures on the calculus at the *Ecole Polytechnique*, which appeared in 1823. The book exhibited his rigour even in its design and printing. This devout Catholic presented his account in *40* lectures, 20 on the differential calculus followed by 20 on the integral calculus—an important religious number, *not* corresponding to the actual number of lectures that he actually delivered at the school [Gilain, 1989, 51–96]. Moreover, each lecture was printed on *exactly* four pages—the theory of limits applied to printing, doubtless intentionally. Until noticing this property, which is not repeated in the reprint in Cauchy's *Oeuvres*, I was puzzled by the concatenation of topics in some lectures. Table 2 is not sufficiently detailed to demonstrate this point, but it shows the range of topics treated.

Building on the notions given in the *Cours d'analyse* (lects. 1–2), Cauchy started with a continuous function and considered the behaviour of its difference quotient $\Delta f(x)/\Delta x$ as the forward difference Δx moved towards zero. Should there be a limiting value, then it was the '*derived function*', written '$f'(x)$'—Lagrange's term and symbol, but in a completely different mathematical context (lect. 3: he also used 'y''). He courted further perplexity by using another traditional calculus word in a new sense (lect. 4). Setting $\Delta x = \alpha h$ with h finite, then if $\Delta f(x)/\alpha$ converged at all, its value was the 'differential' $df(x)$. Thus

$$df(x) = f'(x)\,dx; \tag{3}$$

Table 2. Contents of Cauchy's *Résumé*. Each lecture is exactly four pages in the original printing. A more detailed table is given in [Grattan-Guinness, 1990, 748].

Lectures	**Topics**
	'Warning' (3 pages): aims; status of Taylor's series.
1	Variables, limits, infinitesimals.
2–4	(Dis)continuous functions; derivative and differential.
5	Differentials, including of complex-variable functions.
6–7	Optimae; mean value theorem.
8–9	Partial derivatives and differentials.
10–11	Optimae problems; multipliers.
12–15	Higher-order derivatives and differentials; total differentials.
16–18	(Total) differentials for functions of several variables.
19–20	Polynomial functions; partial fraction expansions.
21–23	Definite integral; evaluations; complex integral.
24–25	Indeterminacy of integrals; 'singular' integral.
26–27	Indefinite integral; definition and properties.
28–31	Integration of some basic functions.
32	'Passage' from indefinite to definite integral.
33–34	Differentiation under the integral sign; double integral.
35	Parametric and successive Differentiation of integrals.
36–38	Taylor's and other series; convergence.
39	Exponential and logarithmic functions.
40	Term-by-term integration of infinite series.
Addition	12 pages: mean value theorems; order of infinitesimals.
Addition	4 pages: Taylor's and MacLaurin series.
Addition	6 pages: partial fraction expansions. [Finally omitted.]

further,

$$\text{if } f(x) = x, \text{ then } dx = h; \ \therefore \ df(x) = f'(x)\,dx, \tag{4}$$

as usual, at least in notation. The mathematics as such is correct; but pity the poor students seeing old symbols with new clothes, especially infinitesimals taking finite values. Nevertheless, this new sense of differential is part of Cauchy's eventual influence [Taylor, 1974]. In this first part of his book he reworked the basic fabric of the differential calculus: partial and total differentials and derivatives, and conditions for optimae including multipliers for constraints.

Cauchy's treatment of the mean value theorem is notable in two respects. Firstly, in the proof of a lemma 'we designate two very small [positive] numbers δ, ε' (lect. 7, (3)), the début of these famous letters. In fact they appear rather rarely in Cauchy's work since he often worked with the associated sequences. Secondly, he also gave the novel extension of the theorem for the quotient of two functions (addn. 1).

The second score of lectures, on the integral calculus, was equally blessed with novelties; some results were elaborated or added in *Exercices* papers of 1826 and 1827. The

integral was a sum, but not of (traditional) infinitesimals; again restricting himself to continuous functions over a range of values $x_0 \leqslant x \leqslant X$, Cauchy selected a finite number of intermediate values $x_1, \ldots, x_{n-1}$, formed the sum $\sum_r (f(x_{r-1})(x_r - x_{r-1}))$, and wondered if it converged to some value as the number of chosen points ever increased. If so, then the limit was written '$\int_{x_0}^{X} f(x)\,dx$' (lect. 21); but while the concept of the integral as limit of a sum was clear, the symbol 'dx' was meaningless. Indeed, it was worse; for he substituted from $(3)_2$ to write '$\int hf(x)$' (after equation (9))!

Cauchy then showed that the limiting value did not depend upon the sequence of partitions used to reach it, or on the choice of x_{r-1} as the values of x for $f(x)$ in the sum (lect. 21). (Soon he will imitate this procedure to define the integral of a function of a complex variable: see §28.7.) He proved the mean value theorem (lect. 26), but not the second theorem, which is due to Ossian Bonnet in mid century; however, he claimed the theorem

$$\text{`}\int_{x_0}^{X} \phi(x) \cdot \chi(x)\,dx = \phi(\xi) \int_{x_0}^{X} \chi(x)\,dx\text{'} \tag{5}$$

for continuous functions $\phi(x)$ and $\chi(x)$, with ξ lying between x_0 and X (lect. 23, (13)). He also laid out the basic theory of double and repeated integrals (lects. 34–35), and evaluated the integrals of various simple functions. He also presented a few properties of the complex integral, but spared the students from the residue calculus (§28.4).

Two theorems were of especial importance. One was the 'fundamental theorem', not Cauchy's name though for the first time a proper theorem as such in his hands: for a continuous and finite-valued function

$$\text{`}d/dx \int_{x_0}^{x} f(x)\,dx = f(x)\text{'} \quad \text{and} \quad \text{`}d/dx \int_{x}^{X} f(x)\,dx = -f(x)\text{'} \tag{6}$$

(lect. 26). Even this success was only half, as he failed to prove the converse relationship, about $\int f'(x)\,dx$.

The other result was the convergence of Taylor's series, whose status was challenged. Instead of grounding Lagrange's algebraic empire, it now suffered the indignity of $\exp(-1/x^2)$, of which all the derivatives were zero at $x = 0$, so that it took *no expansion at all* about that value. Cauchy mentioned this finding at the end of lect. 38, with a remark that the expansion of any other function about $x = 0$ could not be unique; in a profound contemporary paper he gave more examples of such functions and explored the resulting dichotomy between functions and power-series [Cauchy, 1822].

The convergence of Taylor's series now needed examination. By integrating $\int_{x_0}^{x} f(x)\,dx$ successively by parts, Cauchy obtained the integral form of the remainder after n terms and analysed its smallness; he then used the mean value theorem to convert it to the differential form (lects. 35–36). While neither result was new, their role in the issue of convergence was freshly thought out; in particular, the requirement that all preceding derivatives be continuous.

The limitations of the theory in *Cours d'analyse* concerning multiple variables and limits is evident here also. For example, Cauchy called integrals 'singular' if the integrand went to infinite values and/or if the interval was infinite, and in both cases took limiting

values of each limit of the integral without considering simultaneous or successive action: for example, respectively

$$\text{`}\int_{a-\varepsilon}^{a-\varepsilon\mu} f(x)\,dx\text{'} \quad \text{and} \quad \text{`}\int_{-1/\varepsilon}^{-1/\varepsilon\mu} f(x)\,dx\text{'}, \quad \mu > 0 \text{ and constant} \tag{7}$$

(lect. 25, (1)–(4); compare lect. 24, (3)–(4)). Again, he proved that a convergent infinite series of functions was still convergent after integration term by term. The proof depended upon the mean value theorem for integrals, which gave for the integral of the remainder term:

$$\int_{x_0}^{X} r_n\,dx = R_n(X - x_0), \tag{8}$$

where R_n (my symbol) was the value of r_n for some value of x between x_0 and X (lect. 40, (7)). But the smallness of the integral required the uniform smallness of its integrand, which Cauchy did not notice. Similarly, in the discussion of differentiating and integrating under the integral (lect. 33) he unconsciously assumed the integrand to be uniformly continuous on occasion.

5 REACTIONS AT THE *ECOLE POLYTECHNIQUE*: CAUCHY'S LATER BOOKS

In the *Résumé* Cauchy construed

$$\int f(x)\,dx \text{ as the solution of the differential equation } dy = f(x)\,dx \tag{9}$$

(lect. 26, (11); lect. 27, (1)); and in a lecture course at the *Ecole Polytechnique* in 1824 he entertained the second-year students with a wonderful extension of this conception by examining the existence of the solution of

$$dy = f(x, y)\,dx \tag{10}$$

in the same way, taking a partition of values of x and forming the corresponding sum. This is the method now named after him and Rudolf Lipschitz; and the reason why 'Cauchy' does not stand alone is the reaction of the school, who terminated the course as too difficult and forbad completion of the printing of the notes. They disappeared until the late 1970s, when the historian Christian Gilain found 136 printed folios and published them with an excellent introduction [Cauchy, 1981]. I do not discuss them further here, therefore; but the attitude of the school to his teaching requires discussion.

Clearly Cauchy was erecting a beautiful mathematical structure in these two books; but was he meeting student requirements? An event on 12 April 1821 provides the answer. Each lecture period lasted 90 minutes, comprising 30 minutes of discussion of previous material followed by an hour's lecture; but Cauchy seems always to have taken all 90 minutes as his own, and on that day, giving the 65th of the prescribed 50 (sic) lectures on analysis, he continued for another 20 minutes on a new topic; so the students whistled at him and walked out. This was a most serious matter in a military institution, and the

students were sent to barracks; but in an extensive correspondence with the minister, the governor also severely criticised the professor for delivering 'a luxury of analysis no doubt appropriate for papers to be read at the *Institut*; but superabundant for the teaching of students at this school' [Grattan-Guinness, 1990, 713].

This event seems to have been the last straw of dissatisfaction with the professor; yet, apart from the cancellation of the 1824 lecture notes, no change occurred. From 1826 to 1829 graduation examiner de Prony strongly attacked the courses of his former student, but again no major alterations were made [Grattan-Guinness, 1990, 1337–1340]. No evidence survives of the professor's reactions: Cauchy stopped only when his beloved monarch was deposed in the Revolution of July 1830 (in which the students were prominent in the fighting); he abandoned all his posts and stopped the *Exercices,* and joined the royal family into exile as tutor in mathematics of the Bourbon pretender to the throne [Belhoste, 1991, chs. 9–10]. The failure of the school during the 1820s to alter his teaching practice may have been due to the same cause: political life in France had tensioned considerably after the assassination in 1820 of the Duc de Berry, the monarch presumptive and father of the later pretender, and maybe Cauchy's fanaticism for them provided protection.

Further, Cauchy could argue that he had been fulfilling his requirement of providing printed courses. In addition to the cancelled sheets, he published two volumes of lectures on differential geometry [Cauchy, 1826, 1828], again put out by the family firm. The first and main volume (22 lectures, 400 pages) surely went far beyond taught material (or at least, one so hopes), with a detailed analysis of tangents and tangent planes, osculation and curvature, curves in space, and so on, all based upon limits. An outstanding feature was a new theory of orders of contact of curves at a point P, grounded in his theory of infinitesimals (instead of Taylor's series) in terms of the angle between the lines of intersection of each curve with a circle with centre P (lect. 9). The second volume (4 lectures, 123 pages) dealt with integration matters such as rectification of curves, quadrature of surfaces, and cubature of solids.

Again, with the *Résumé* apparently sold out, Cauchy planned to write two much larger volumes on the calculus, though only the one on the differential calculus appeared [Cauchy, 1829] before his roof fell in. At 289 pages it was over four times longer than its predecessor, but mostly considerable elaborations of the *Résumé*, especially concerning functions of several real variables and of a complex variable; much of the new material came in an addition, on methods of approximating to the roots of an equation by using truncated Taylor series.

So the professor fulfilled much of his obligations, at least as he saw them. However, none of his four books ever entered the list of textbooks recommended to the students of the school.

6 THE GRADUAL INFLUENCE OF CAUCHY'S DOCTRINE

When Cauchy left Paris he was stripped of his chair at the *Ecole Polytechnique.* His courses were taken over at first by his assistant G.G. Coriolis, and his post by C.L.M.H. Navier. Both men were engineers, and did not develop many of Cauchy's main notions. But after Navier's death in 1836 the post went to J.M.C. Duhamel and from 1840 to Charles Sturm,

who were more sympathetic to them. In particular, after its posthumous publication in the late 1850s Sturm's own *Cours d'analyse* was very influential, with editions until 1929.

Cauchy himself followed the royal family around Europe, especially Italy; and he gave a short version of the *Cours d'analyse* in Turin, which was published there as [Cauchy, 1833]. He gained some Italian followers, and some of his material appeared in Italian. Upon his return to France in 1838 he refused to sign an oath of allegiance to the ruling administration and so could not take up any positions; but such obligations were removed after the 1848 revolution, and from 1849 until his death in 1857 he taught mathematical astronomy at the *Faculté des Sciences.* Publicity for his doctrine, therefore, passed to others' hands: in particular, between 1840 and 1868 fellow Catholic and his former student at the *Faculté* the abbot Moigno (1804–1884) produced several volumes on the calculus, and also on the calculus of variations and parts of mechanics, based upon his previous teaching.

So here are some markers for the adoption of Cauchy's doctrine; but the other traditions maintained good positions, especially the Leibniz–Euler differential and integral version with the differential coefficient (§14). A good example is Britain, for it had maintained Newton's fluxional calculus until the 1800s, when switches were made especially to Lagrange's algebraic approach, and also to some extent to Leibniz and Euler. The further move to Cauchy was fitful [Rice, 2001]. An important author is Augustus de Morgan (1806–1871), who outlined the theory of limits in his *Elements of algebra* (1835), and treated mathematical analysis in a mammoth *Differential and integral calculus* (1842). He made some acknowledgement to Cauchy, but not very often; and while he handled continuous functions and the calculus in broadly Cauchyesque terms, he also included a long chapter on divergent infinite series. Again, William Whewell argued for the merits of limits, but he had in mind his Trinity College predecessor Isaac Newton more than any Frenchman.

An important case for Cauchy's doctrine is Germany (then, the German states). As is shown at the head of this article, the *Cours d'analyse* was translated twice, the first one quite early in 1828; the *Résumé* was not translated, but Cauchy's doctrine was adopted in that decade, initially by Martin Ohm in Berlin. Further, Cauchy's later books mentioned in Section 5 appeared in the 1840s thanks to C.H. Schnuse, who translated many other French mathematical books including some of the Moigno material. Some young German-speaking mathematicians applied the doctrine notably; already in the 1820s J.P.G. Lejeune-Dirichlet and the Norwegian N.H. Abel. However, as elsewhere the other traditions of the calculus remained widely taught. Dirichlet's work inspired Bernhard Riemann's contributions to mathematical analysis (§38.2).

The most striking case is Karl Weierstrass (1815–1897). From the late 1850s he was professor at Berlin, and arguably the leading mathematician of the world. The influence of his 30 years' lecturing is immense (although, as was mentioned in §0.3.2, he never published his courses, so that there is no one text which qualifies for an article in this book). For complex-variable analysis he introduced a new foundation based upon power-series expansions, and so was in competition with Cauchy; but in real-variable analysis he and his students not only adopted all parts of Cauchy's doctrine but also refined them in various ways: for example, *multiple* limits and the consequences for continuity and convergence, existence theorems and definitions of irrational numbers [Grattan-Guinness, 1980]. He also popularised the use of Cauchy's 'ε' and 'δ'. So Cauchy was super-vindicated, as it were;

but the origins of Weierstrass's approach lie buried in his obscure school-master days of the 1840s and 1850s [Dugac, 1973] and relate much to elliptic functions, a topic on which—rarely—Cauchy published little.

A curious case is Russia. As we see at the head of this article, some of the *Résumé* was available already in 1831 thanks to B.Ya. Bunyakovsky, and the *Cours d'analyse* was translated in 1864 (but published in Leipzig). Before the 1870s Russia was not a significant country for mathematics; but Bunyakovsky and especially M. Ostrodgradsky had contributed notably to aspects of mathematical analysis, and Cauchy's doctrine played a role [Yushkevich, 1968, ch. 14].

It seems clear that the ultimate supervention of Cauchy's doctrine over its competitors, especially for those mathematicians concerned with rigour and proof, was inspired above all by the *critical spirit* which he had elevated far above the levels available in other traditions: systematically if–then mathematics, sufficient and/or necessary conditions for the truths of theorems, and especially the fundamental theorem of the calculus at last a pukka theorem, even if he only started half of it in (4). But the "victory" was a gradual one—a fascinating but little-studied international story towards which this section supplies only notes.

BIBLIOGRAPHY

Belhoste, B. 1991. *Augustin-Louis Cauchy. A biography,* Berlin: Springer.

Bottazzini, U. 1986. *The higher calculus . . .*, Heidelberg: Springer.

Bottazzini, U. 1992. 'Geometrical rigour and "modern analysis"; an introduction to Cauchy's *Cours d'analyse*', in Cauchy, *Cours d'analyse* reprint, Bologna: CLUEB, xi–clxvii.

Cauchy, A.-L. *Works*. *Oeuvres complètes,* 2nd series, Paris: Gauthiers–Villars, 1905–1974.

Cauchy, A.-L. 1822. 'Sur le développement des fonctions en séries . . .', *Bulletin des sciences par la Société Philomatique de Paris,* 49–54. [Repr. in *Works*, vol. 2, 276–282.]

Cauchy, A.-L. 1826, 1828. *Leçons sur les applications du calcul infinitésimal à la géométrie,* 2 vols., Paris: de Bure. [Repr. as *Works*, vol. 5.]

Cauchy, A.-L. 1829. *Leçons sur le calcul différentiel,* Paris: de Bure. [Repr. in *Works*, vol. 4, 263–609.]

Cauchy, A.-L. 1833. *Résumés analytiques,* Turin: Imprimerie Royale. [Repr. in *Works*, vol. 10, 5–184.]

Cauchy, A.-L. 1981. *Equations différentielles ordinaires. Cours inédit (fragment)* (ed. C. Gilain), Paris: Etudes Vivantes; New York: Johnson.

Dugac, P. 1973. 'Eléments d'analyse de Karl Weierstrass', *Archive for history of exact sciences, 10,* 41–176.

Gilain, C. 1989. 'Augustin-Louis Cauchy (1789–1857)', *Bulletin de la Société des Amis de la Bibliothèque de l'Ecole Polytechnique,* no. 5, 145 pp.

Grabiner, J.V. 1981. *The origins of Cauchy's rigorous calculus,* Cambridge, MA: MIT Press.

Grattan-Guinness, I. 1970. *The development of the foundations of mathematical analysis from Euler to Riemann,* Cambridge, MA: MIT Press.

Grattan-Guinness, I. 1980. 'The emergence of mathematical analysis and its foundational progress', in (ed.), *From the calculus to set theory, 1630–1910: an introductory history,* London: Duckworth, 94–148. [Book repr. Princeton: Princeton University Press, 2000.]

Grattan-Guinness, I. 1990. *Convolutions in French mathematics, 1800–1840. From the calculus and mechanics to mathematical analysis and mathematical physics,* 3 vols., Basel: Birkhäuser and Berlin (DDR): Deutscher Verlag der Wissenschaften.

Laugwitz, D. 1987. 'Infinitely small quantities in Cauchy's textbooks', *Historia mathematica, 14,* 258–274.

Pringsheim, A. and Faber, G. 1908. 'Algebraische Analysis', in *Encyklopädie der mathematischen Wissenschaften,* vol. 2, sec. 3, 1–46 (article II C1).

Rice, A.C. 2001. 'A gradual innovation: the introduction of Cauchyian calculus into mid-nineteenth-century Britain', in *Proceedings of the Canadian Society for the History of Mathematics, 13,* 48–63.

Taylor, A.E. 1974. 'The differential: nineteenth and twentieth century developments', *Archive for history of exact sciences, 12,* 355–383.

Valson, C.A. 1868. *La vie et les travaux du Baron Cauchy,* 2 vols., Paris: Gauthiers–Villars. [Repr. Paris: Blanchard, 1970.]

Yushkevich, A.P. 1968. *Istoriya mathematiki v Rossii do 1917 goda,* Moscow: Nauka.

CHAPTER 26

JOSEPH FOURIER, *THÉORIE ANALYTIQUE DE LA CHALEUR* (1822)

I. Grattan-Guinness

This book contains the first extended mathematical account of heat diffusion, itself the first major mathematicisation of a branch of physics outside mechanics. The mathematical importance lay mainly in Fourier series and integrals.

First publication. Paris: Firmin Didot, 1822. xxii+639 pages.

Photoreprints. Breslau: Köbner, 1883. Paris: Gabay, 1988.

Reprint. As *Oeuvres*, volume 1 (ed. G. Darboux), Paris: Gauthier–Villars, 1888.

English translation. The analytical theory of heat (trans. A. Freeman), Cambridge: Cambridge University Press, 1878. [Photorepr. New York: Dover, 1955.]

German translation. Die analytische Theorie der Wärme (trans. B. Weinstein), Berlin: J. Springer, 1884.

Spanish translation. Teoría analatica del calor (trans. D. Redondo Alvarado), Madrid: Universidad Politécnica, 1992.

Other translations. Japanese (trans. H. Yoshida from the English translation), Tokyo: 1993. Chinese (trans. Z. Gui), Hefe: 1994.

Related articles: Cauchy on real-variable analysis (§25), Riemann on trigonometric series (§38), Baire and Lebesgue (§59), Bochner (§74).

1 EDUCATION AND EMPLOYMENTS

Somewhat unusually for a mathematician, Jean Baptiste Joseph Fourier (1768–1830) followed an eventful career outside his intellectual activities [Herivel, 1975, pt. 1; Dhombres and Robert, 1998, chs. 2–6]. The initial cause was the French Revolution of 1789. At the

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

time he was living in his home town of Auxerre; he seems to have conducted himself honourably, and so was arrested in 1794 though soon released. He was soon nominated by a neighbouring town to be its student at the *Ecole Normale* in Paris, newly established for the training of teachers. Despite employing eminent professors such as Joseph Louis Lagrange (1736–1813), Pierre Simon Laplace (1749–1827) and Gaspard Monge (1846–1818), poor planning and insufficient funding led to its closure after four months. However, he had made enough impression to be appointed as a junior teacher at the *Ecole Polytechnique*, another new institution with which Lagrange, Laplace and Monge were also involved, but one that was to endure.

Fourier doubtless hope to emulate his seniors in his career; but in 1798 he was chosen, seemingly by Monge, to join Bonaparte's expedition to Egypt. Fourier played a prominent role in the civil and academic sides of the occupation; in particular, he led one of scientific teams to examine the ancient ruins and artefacts. He stayed until the defeat by the British in 1801, when he returned to Paris and the *Ecole Polytechnique*. However, Bonaparte saw better uses for the administratively gifted, and in 1802 appointed him Prefect of the *département* of Isère, based at Grenoble.

The region was backward, on the border with Italy. Fourier served it until 1815, with great energy and distinction. He revitalised both education and industry, and launched projects such as the draining of a large area of marshland. Yet he also found time to work on the vast report on the studies of Egypt, helped by some periods in Paris (four days away by carriage). His main contribution to the multi-volume *Description de l'Egypte* was a preface, widely admired for its style upon its publication in 1809.

2 THE CHRONOLOGY OF FOURIER'S RESEARCHES

In addition, somehow during this period Fourier also made most of his main scientific contributions, many of them quickly. He never stated the precise stimulus to mathematicise heat diffusion; apart from the novelty, and maybe experiencing Egyptian heat followed by French Alpine cold, a source could have been a pioneering but shaky short paper published in 1804 by Laplace's follower J.B. Biot. At all events, Fourier had written a substantial manuscript by 1805, and presented an enormous one to the scientific class of the *Institut de France* in December 1807. It contained the diffusion equation for several solid bodies of finite dimension, solutions by various forms of Fourier series (to use the modern name) and by the function we now call the Bessel function $J_0(x)$, and some experimental results [Fourier, 1807].

Laplace and Lagrange were among the examiners, and they reacted sceptically for different reasons, explained in sections 3 and 4 below. (Monge was also a member, and presumably welcomed the paper.) Nevertheless, a prize problem was proposed by the class for 1811, and Fourier submitted a still longer piece [Fourier, 1811] which rehearsed the previous one but also contained the solution by Fourier integrals of the diffusion equation for infinite solid bodies, and some physical aspects of heat (apparently motivated by discussions with Laplace). He won the prize, but the report of the commission (which again included Laplace and Lagrange) was critical of some aspects; thus publication in the journal for papers submitted to the class by *savants étrangers* seemed distant, especially as it had not appeared since 1806 anyway. So he concentrated on a third version, as a book.

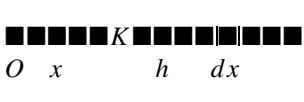
But politics came in to Fourier's life again, this time over the abdication and then return of Emperor Napoléon in 1814–1815. Fourier's position was especially tricky because of his Prefecture; in 1815 he accepted Napoléon's "promotion" to the prefecture of the *département* of Yonne based at Lyon, but he resigned in protest over policy before Napoléon's Hundred Days were over. He moved to Paris, without a post. But now contacts from the *Ecole Polytechnique* helped: a former student, now Prefect of the *département* of the Seine, appointed him head of the Bureau of Statistics. Fourier improved his personal situation quickly; he was elected to the restored *Académie des Sciences* in 1817, and five years later even to the influential post of a *secrétaire perpétuel*. His book also appeared then. He also managed to get his 1811 paper published in the *Mémoires* of the *Académie* in 1824 and 1826, although he had not been a member when he won the prize. The 1807 paper remained unpublished; an edition of it is included in [Grattan-Guinness and Ravetz, 1972], along with parts of the 1805 predecessor.

In 1816 Fourier published a paper announcing the imminent appearance of a book on both the mathematical and the physical aspects of heat [Fourier, 1816]; but six years were to pass before a book was published, and it covered only the mathematical sides. In the 'preliminary discourse' he stated that its writing and printing had taken a long time (p. xvii). Table 1 summarises not only the book but also comparable passages in the two main earlier papers; all three sources are cited below by article numbers.

3 HEAT DIFFUSION, INTERNAL AND SURFACE

In somewhat tedious detail, Fourier began by exploring the known properties and parameters for the study of heat diffusion. There were three main ones of the latter: conduction internally within a solid body ('*K*') and externally through its surface or boundary into the environment ('*h*'); and specific heat ('*C*') (arts. 26–39). Assuming them to be constant, he used them to define quantity of heat at a point or section of the body. He took the flow of heat to be uniform, and temperature change linear with respect to distance. He drew upon Newton's law of cooling, that the flow of heat through a domain was proportional to the temperature difference across it (arts. 64–68; 429, no. 3). But he became aware of its fallibility and altered parts of arts. 31–38 of the book while on proof; the original text is found in the copy in the *Bibliothèque Nationale*, Paris [Grattan-Guinness, 1990, 1330–1332].

To mathematicise the phenomenon Fourier used the standard differential and integral calculus in the version developed by Euler with the differential coefficient (§14). In his first example a straight bar ('prism') of square cross-section with (tiny) side $2l$ diffused both internally and externally; when in thermal equilibrium its temperature at point x distant from end O was v, with the environment set at temperature 0. At the infinitesimal slice with face x and constant thickness dx, $8lh\,dx.(v - 0)$ entered into the environment, while



$$4l^2K\left[d(dv/dx) + dv/dx\right] - 4l^2K\,dv/dx, \quad \text{or} \quad 4l^2K\,d(dv/dx) \tag{1}$$

Table 1. Contents of Fourier's book, compared with the 1807 and 1811 papers.
The numbers are of articles: 91–94 of the book were mis-numbered 84–87, but are corrected here, as in all later editions. Since Fourier made various small and large changes in text, the comparisons are somewhat gross, especially for arts. 1–80. A blank indicates that there is no corresponding text.

1807	**1811**	**Book**	**Topics**
	1	pp. i–xxii	'Preliminary discourse': physical properties of heat; mathematics.
Preface	3, 1	1–21	Aims, results; heat constants and properties.
15–16	2	22–56	Conductivity; specific heat; action of heat.
17–18	4–6	57–72	Communication and linearity of heat diffusion.
19–22	7–8	73–80	Steady-state diffusion in bar.
		81–91	Heating in closed spaces.
		92–100A	Diffusion in three dimensions; at a point.
23–24	9–10	101–110	Diffusion equation for ring; special solution.
25–28	11–14	111–131	Equations for sphere, cylinder, infinite bar, cube.
29–31	15	131–162	General diffusion equation; surface diffusion; special cases.
32–47	16–20	163–189	Diffusion equation for lamina; Fourier series; infinite "matrix".
48–49	21, 30	190–206	Solution for lamina.
50–84	21–29	207–237	Series for general function; properties.
76–94	31–37	238–236	Diffusion in ring.
1–13, 95–96	38–43	247–282	n-body model.
97–114	[44]–50	283–305	Diffusion in sphere; surface diffusion; non-harmonic series.
116–139	51–56	306–320	Diffusion in cylinder; $J_0(x)$.
140–151	57–61	321–332	Steady-state diffusion in infinite bar; multiple series.
152–158	62–65	333–341	Diffusion in cube; multiple series.
	66–72	342–364	Diffusion in infinite body; Fourier integrals.
	73–79	364–385	Laplace's solution of the diffusion equation.
		386–395	Highest temperature in infinite body.
		396–433	Fourier integral and other solutions to differential equations.

passed through the slice from its neighbour. Thermal equilibrium ensured that the two quantities were equal; so the diffusion equation took the form

$$8lh\,dx(v-0) = 4l^2K\,d(dv/dx), \quad \text{or} \quad d^2v/dx^2 = (2h/Kl)v \tag{2}$$

(arts. 74–75). In a novel move Fourier also derived $(2)_2$ by considering the quantity of heat in the bar from O to x, thereby obtaining an integro-differential equation and differentiating it with respect to x (art. 73). The interest was in using only one slice instead of two.

In the dynamic situation over time, the difference between the two quantities in $(2)_1$ during a constant infinitesimal interval dt would balance the rise in temperature dv, so that

$$4l^2 K\, d(dv/dx)\, dt - 8lh\, dx.v\, dt = CD4l^2\, dx\, dv. \tag{3}$$

$$\therefore\ K d^2v/dx^2 - (2h/l)v = CD\, dv/dt, \tag{4}$$

where D was the (constant) density of the body, and the ratios of differentials were understood as *partial* differential coefficients of v as a function of both x and of t. Curiously, Fourier did not show (4), but he found its analogue for the ring, where x was the angular variable (arts. 102–105).

These two cases applied to bodies of one dimension, when both coefficients of conductivity obtained. For continuous solid bodies such as the cube, Fourier showed that interior conduction in three spatial variables was expressed by the equation

$$`dv/dt = (K/CD)(d^2v/dx^2 + d^2v/dy^2 + d^2v/dz^2)' \tag{5}$$

(arts. 126–128). The distinctive feature of (4) and (5) among the partial differential equations then known was the second-order derivation with respect to x and the first-order for t. In steady state it became Laplace's equation (as we now call it) (arts. 121–123).

Surface diffusion required a separate equation, which Fourier found by equating the internal flow adjacent to the surface with the external flow from it into the environment; for example, for the cube of side $2l$ centred at the origin O flow over time dt through and out of a surface infinitesimal rectangle perpendicular to Ox with sides dy and dz gave

$$-K(dy\, dz)(dv/dx)\, dt = h(dy\, dz)(v - 0)\, dt; \tag{6}$$

$$\therefore\ hv + K\, dv/dz = 0 \quad \text{when } x = \pm l \tag{7}$$

(art. 129). For the general (smooth) surface he showed that at a point with direction ratios (m, n, p)

$$K(m\, dv/dx + n\, dv/dy + p\, dv/dz) + h\sqrt{m^2 + n^2 + p^2}\, v = 0 \tag{8}$$

(arts. 146–154). Mathematically such equations had already appeared in hydrodynamics; but this one represented physical flow, and in itself was a source of influence (section 8 below).

Neither in his analysis nor in the general discussion did Fourier make any commitments as to the nature of heat; he took it just as a phenomenon, with cold as its opposite. Here he was in accord with the Swiss physicist Pierre Prevost, whom he visited in Geneva (not far from Grenoble) in 1804 and also corresponded [Weiss, 1988]. However, in 1807 this stance disappointed Laplace, who had recently launched an ambitious programme of molecular mathematical physics (§18.8); in 1810 he re-derived the diffusion equation in this way, construing heat as central action between molecules which was known only to decline rapidly with distance from source, and cumulative action expressed as an integral involving it. Fourier neither affirmed nor rejected this method.

4 FOURIER SERIES AND THEIR FUNCTION

The mathematical task for Fourier was to solve the diffusion equation (2) for one-dimensional bodies, or some version of the internal equation (5) for solid bodies with (7) serving as a boundary condition. In both cases the initial temperature distribution function was also used.

Fourier always solved (5) by the method of separating the variables, leading to an ordinary differential equation in a spatial variable. To that end he used infinite trigonometric series, producing solutions such as

$$v = \sum_{r=0}^{\infty} (a_r \cos rx + b_r \sin rx) \exp -Kr^2 t. \tag{9}$$

The initial data $v = f(x)$ when $t = 0$ led in (9) to

$$f(x) = \sum_{r=0}^{\infty} (a_r \sin rx + b_r \cos rx), \tag{10}$$

with the coefficients given by

$$\pi a_r = \int_{-\pi}^{\pi} f(x) \cos rx \, dx \quad \text{when } r \neq 0, \quad 2\pi a_0 = \int_{-\pi}^{\pi} f(x) \, dx, \tag{11}$$

and a formula corresponding to $(11)_1$ for the b_r. All these results were known by 1807: the discussion in the book was long (arts. 163–237), and the formulae turned up later for the special bodies.

I have used our familiar symbols for summation, subscripts and definite integrals; and Fourier is an important source of their growth in popularity. He used subscripts much more systematically than normal at that time; and he *invented* the notation '$\int_a^b$' for the definite integral as an elaboration of Leibniz's symbol (arts. 222, 231; first published in his paper [1816]).

Two major issues about the series are discussed now. The first concerns *representability*. (10) had been known before Fourier, especially to Euler and Lagrange, but they had rarely been advocated as a general solution to a differential equation; much preferred was the functional form, where f appeared explicitly rather than buried in integrals (11), or if necessary that by infinite power series. When Fourier put them forward in 1807 Lagrange objected that (10) could not be general since the series were periodic; further, the sine and cosine series were also respectively odd and even. Fourier had already dealt well with feature in the manuscript, and repeated it in notes for Lagrange: for all three kinds of series '=' in (9) and (10) pertained only over the interval specified for the physical problem, over which the integrals were defined; outside it function and series (usually) parted company. He even found three different series for the function $x/2$ over $0 \leqslant x \leqslant \pi$, and drew them over several periods (1807, art. 68). This geometrical of thinking, typical of Fourier and maybe showing influence from Monge, lay outside the algebraic realm of Lagrange, who

never accepted the answer. While the same examples were given in the 1811 paper and the book (art. 225) Fourier omitted most of the diagrams, maybe to make the point less clear and so more acceptable.

The point is important, for without it further questions about the series need not be asked. For example, it had affected Daniel Bernoulli's advocacy of the series (on physical grounds only) in the famous debate in the mid 18th century on the vibrating string problem (§59.1), where others had preferred the functional solution. Fourier knew that background, and commented rather briefly upon it (art. 230).

The second issue is *generality*. In order to compete with other methods, Fourier had to show that discontinuous functions could be represented. This was duly done, though with the jumps joined by vertical lines in the diagrams in the 1807 paper, thus making the functions contiguous (an example appears in the book at art. 232). But he did not envision the 'Gibbs phenomenon', where the vertical line has to be extended a little on both sides of the jump [Hewitt and Hewitt, 1979]: it arises from distinguishing repeated from multiple limits, which nobody in Fourier's time fully grasped (compare §25.3).

The best solution to the problem of generality would be a proof of their convergence, and to the function. Fourier offered one in his book (arts. 415–416; 279 on the convergence of the full solution (9)); but it was defective in claiming that "any" function $f(x)$ should satisfy $(10)_2$. Initial resolution would come from other hands near the end of his life (section 7).

5 CALCULATING AND INTERPRETING THE COEFFICIENTS

Fourier gave two methods. The second was the (now) usual one of multiplying through $(10)_2$ by $\cos rx$ and integrating between $-\pi$ and π (arts. 223–224 for sine and cosine series); in a later paper Fourier thanked S.F. Lacroix (who had also been a member of his 1807 jury) for pointing out that Euler had used it before him. But his first method was quite different: $f(x)$ and the trigonometric functions were expanded in their Taylor series and the coefficients of powers set equal. An infinitude of linear equations resulted; the solution $(11)_1$ was finally obtained after a monstrous bag of clever tricks with finite and infinite series and products.

This is the pioneer effort in the theory of infinite matrices [Bernkopf, 1968]; but why was it preserved in the 1811 paper and the book (arts. 207–221; 171–177 for the function $f(x) = 1$) when the much more convenient alternative was also available? Partly, no doubt, Fourier did not want to lose some interesting mathematics. However, he had also spotted a defect with the other method: it worked even if terms were missing from the series, whence, if the corresponding coefficients were not zero, the resulting series was incorrect [Grattan-Guinness and Ravetz, 1972, 237–239]. This feature was not to be noticed again until the 1890s onwards, when it was resolved in terms of completeness of function spaces: Fourier noted it briefly in his book (arts. 424–425), where he also gave the so-called 'Parseval formula' (art. 235, no. 2) without however noting any connection between the two points.

This first method may also have been preserved because of a feature of the physics. In his first attempt to mathematicise heat diffusion Fourier had used the known method of discrete modelling: for example, the continuous ring was replaced by n equal separate mass

set in the corresponding circle and exchanging heat and cold. He found a set of simultaneous ordinary differential equations, which he solved by separation of variables, leading to *finite* trigonometric series in the angular variable (1807, arts. 1–13). The continuous case was found by letting $n \to \infty$; but he found the false solution of steady temperature, and so started again with the differential modelling [Grattan-Guinness and Ravetz, 1972, 81–82]. This analysis was preserved in the 1811 paper and the book (arts. 247–282), even though it had become nostalgia. The cause of error was a mis-specification of K: perhaps because of this mishap, in the later versions he indicated the units and dimensions of all his physical parameters (art. 161), pioneering the dimensional analysis of scientific theories [Macagno, 1971]. In his empiricist spirit mentioned in section 3, Fourier did not use either (10) or this finite predecessor to argue that heat was a waval phenomenon, even though this view was gaining some adherents at that time [Brush, 1976, ch. 9].

A related question was the definability of the integrals in (11). Fourier construed them geometrically as areas (arts. 220, 229, 415–417) rather than the algebraic Lagrangian inverse of the derivative, but he did not analyse area itself (compare A.-L. Cauchy in §25.4).

6 NON-HARMONIC SERIES, AND THE BESSEL FUNCTION

Fourier had to make two important modifications to his solutions. One concerned solid bodies: the coefficients in the trigonometric terms were themselves unknown, to be determined from the surface condition (7). They turned out to be the roots n_r of transcendental equations in n, such as

$$\tan nX = nX(1 - hX) \tag{12}$$

for the sphere of radius X (arts. 284–288; compare 328 for the infinite bar). The associated solution, later called 'non-harmonic', was

$$v = \sum_{r=0}^{\infty} (a_r \cos n_r x + b_r \sin n_r x) \exp -K n_r^2 t. \tag{13}$$

It was essential for the physics to show that (12) possessed only real roots; for otherwise the exponential time terms in (13) would not decay as $t \to \infty$. Luckily Fourier was in familiar territory, for already in his Auxerre days he had generalised Descartes's rule of signs into an upper bound for the number of roots of a polynomial equation lying within a given interval of values [Grattan-Guinness and Ravetz, 1972, 8–12]. Here he used geometrical illustrations to show that (12) and its kin possessed an infinitude of real roots; but the banishment of complex roots was not completely secured.

The other main change occurred when Fourier analysed diffusion in the cylinder. Casting the diffusion equation in cylindrical polar co-ordinates, separation of variables led to the differential equation for the function $u(x)$ of the axial variable x

$$d^2u/dx^2 + (1/x)\, du/dx + (mCD/K)u = 0, \tag{14}$$

with m to be determined by consideration of surface diffusion (7). This time the series did *not* work; so he followed the tradition of solution by power-series in x, and after a virtuoso

manipulation of analytical techniques he found many of the basic properties of the function $J_0(x)$ which we now name after F.W. Bessel: not only the series expansion but also its generating function the integral form, orthogonality, the expansion of "any" function in a series of such functions, and a good though again not conclusive argument for the reality of the associated roots. Again this was done by 1807; it received a substantial statement in the book (arts. 306–319), together with emphasis on series expansions of functions in general (art. 424; 428, no. 6).

7 THE LAPLACE AND FOURIER INTEGRALS

Not achieved by 1807 was the analysis of diffusion in an infinite body. Series would not do, not only in view of their finite periodicity but especially because the methodology of surface diffusion embodied in (7) failed, since that kind of body had no such surface.

As part of his response to Fourier's paper, Laplace showed in 1809 that the solution of the diffusion equation (4) in such circumstances was proved by an *integral*; in the notation of (9),

$$\sqrt{\pi}v = \int_0^\infty f\left(x + 2s\sqrt{t}\right) \exp\left(-s^2\right) ds, \quad \text{where } f(x) = v(x, 0). \tag{15}$$

This solution is not to be confused with the (rather mis-named) 'Laplace transform'.

With this big hint, Fourier found in time for the 1811 prize paper his own integral solution. The initial conditions yielded the integral formula now named after him, in versions such as

$$2\pi f(x) = \int_{-\infty}^{\infty} f(u)\, du \int_{-\infty}^{\infty} \cos\left(q(x-u)\right) dq. \tag{16}$$

His proof relied upon rather hair-raising tricks with infinitesimals to convert sums into integrals (arts. 344–346). He also gave versions of (16) over positive and negative values of x; and, as with the series, he gave sine and cosine forms. To some extent he distinguished multiple from repeated integrals. Both solutions, especially his own, were discussed at length in the book (arts. 348–384, 396–413). Noting the similarity of form between the kernels of the diffusion equation and of his integrals, he also obtained integral solutions to more general linear partial differential equations with constant coefficients by using differential operators (arts. 401–404), an approach which had gained some currency. However, Cauchy rejected such methods, replacing the solutions with forms using complex-variable Fourier integrals; he had found (16) for himself in 1817, and his use of the integrals added much to their early prestige.

8 LATER WORK AND RECOGNITION

As well as organising the publication of his 1811 paper and book when based in Paris, Fourier published some papers on both mathematical and physical aspects of heat theory. One notable rise in prestige came in 1820 when he applied Fourier integrals to the cooling of a large sphere to estimate the age of the Earth. Laplace found another value from an analysis using spherical harmonics, but praised Fourier's contribution.

This was pointed approbation, as Laplace's faithful follower S.-D. Poisson (1781–1840) was then reworking much of Fourier's physics in Laplacian molecularist manner. He solved the diffusion equation either by Laplace's (15) or in Lagrangian style by taking the terms of Fourier's series as coefficients of a power series, forming the integral now named after him [Grattan-Guinness, 1990, ch. 12]. Most of his cases were desimplifications of some of Fourier's, such as diffusion in a bar made in two parts of different materials, and of the sphere into an environment with non-constant temperature. His best contribution was an excellent argument of 1826 for the reality of the roots of a class of transcendental equations including (12).

Fourier also wrote on radiant heat and its motion within liquids, and in 1823 he conducted experiments on thermo-electric effects with H.C. Ørsted in the latter's visit to Paris. Later he declared the intention of writing a separate book on the physical aspects of heat theory, presumably including the experiments on heat diffusion which he had reported in his 1807 and 1811 papers [Grattan-Guinness and Ravetz, 1972, ch. 20]; but no manuscript of it exists in his *Nachlass*.

Fourier also planned a book on the theory of equations and related topics (including what we recognise now as linear programming), and the publishable material appeared posthumously by his follower C.L.M.H. Navier [Fourier, 1831]. He was attracting the interest of several members of the new generation of mathematicians, with 1829 a particularly good year. A Swiss immigrant, Charles Sturm, improved to an equality his old theorem on the upper bound of roots of a polynomial within a given interval of values (section 6). Auguste Comte began his first major set of lectures advocating 'positive philosophy', much inspired by Fourier's philosophical stance over heat; the published version (1836–1842) was dedicated to him. Finally, a recent German visitor, J.P.G. Lejeune-Dirichlet, partly influenced also by Cauchy's recent improvement of rigour in mathematical analysis (§25), produced a classic proof of the convergence of Fourier series under the sufficient conditions that the function possess only a finite number of turning values, discontinuities and (later) points of infinitude (§38.2).

Dirichlet also proved one of Fourier's stated solutions of a diffusion problem; and others to became interested in that subject included the Russian visitor Michel Ostrogradsky, and compatriots J.M.C. Duhamel and Gabriel Lamé [Bachelard, 1928]. The analyses usually deployed the series and/or integral solutions, which were appearing also in other applications; for example, with Navier in elasticity theory.

During the 1820s Fourier also offered some striking thoughts on statistics in connection with his directorship of the Bureau. Among other activities, he helped secure a new chair in 'Egyptology' at the *Collège de France* for his protégé from Grenoble days, Jean Champollion. By the time of his death in 1830 this non-standard and innovative scientist had become establishment.

9 ON THE LATER IMPACT

Fourier's book became a standard source for both students of heat diffusion and of Fourier series and integrals, and solution of linear differential equations in some generality [Burkhardt, 1908]. It began to be used abroad: a particularly notable example is William

Thomson; when aged 16 years in 1840 he took the book as holiday reading, and was soon writing papers on it. Responding warmly to Fourier's positivism (as Comte had called it), he was inspired by the handling of diffusion to envision flow as a major notion in mathematical physics (compare §40). An important stimulus for the series was suggested by G.S. Ohm in 1843: refining Bernoulli on the vibrating string problem, he took the terms to denote super-harmonics in acoustics, a move picked up later by Hermann von Helmholtz, Lord Rayleigh, and many others (§45). On the pure side, Dirichlet's proof led to an important stream of researches, especially Bernhard Riemann's essay of 1854 on trigonometric series (§38) and then on to set theory and measure theory (§59) [Paplauskas, 1966].

By the 1860s Fourier's book was part of the furniture. As shown at the head of this article, in the 1870s and 1880s it received two translations, a photoreprint, and a reprinting as the first of the two volumes of an edition of his scientific writings. Only George Green's masterpiece of 1828 on potential theory (§30) matches it for later re-issue of a publication of that time. Indeed, both books have enjoyed reprints in modern times.

BIBLIOGRAPHY

Bachelard, G. 1928. *Etude sur l'évolution d'un problème de physique. La propagation thermique dans les solides*, Paris: Vrin. [Repr. 1973.]

Bernkopf, M. 1968. 'A history of infinite matrices', *Archive for history of exact sciences, 4*, 308–358.

Brush, S.G. 1976. *The kind of motion we call heat*, 2 vols., Amsterdam: North-Holland.

Burkhardt, H.K.F.L. 1908. 'Entwicklungen nach oscillirenden Functionen und Integration der Differentialgleichungen der mathematischen Physik', *Jahresbericht der Deutschen Mathematiker-Vereinigung, 10*, pt. 2, xii+1804 pp.

Dhombres, J. and Robert, J.-D. 1998. *Fourier. Créateur de la physique–mathématique*, Paris: Belin.

Fourier, J.B.J. 1807. 'Mémoire sur la propagation de la chaleur', manuscript in *Ecole Nationale des Ponts et Chaussées*, Paris, ms. 1851; published in [Grattan-Guinness and Ravetz, 1972]. [Submitted to the *Institut* in 1807.]

Fourier, J.B.J. 1811. 'Théorie du mouvement de la chaleur dans les corps solides', *Mémoires de l'Académie Royale des Sciences, 4* (1819–20: publ. 1824), 185–555; 5 (1821–22: publ. 1826), 153–246. [Second part repr. in *Oeuvres*, vol. 2, 1–94. Submitted to the *Institut* in 1811. Manuscript in *Académie des Sciences*.]

Fourier, J.B.J. 1816. 'Théorie de la chaleur', *Annales de chimie et de physique, (2) 3*, 350–375. [Authorship attributed. Most regrettably, not in his *Oeuvres*.]

Fourier, J.B.J. 1831. *Analyse des équations indéterminées* (ed. C.L.M.H. Navier), Paris: Firmin Didot. [German transs.: Braunschweig: Meyer, 1846; Leipzig: Englesmann, 1902 (*Ostwalds Klassiker der exakten Wissenschaften*, no. 127).]

Grattan-Guinness, I. 1990. *Convolutions in French mathematics, 1800–1840*, 3 vols., Basel: Birkhäuser; Berlin: Deutscher Verlag der Wissenschaften.

Grattan-Guinness, I. with the collaboration of Ravetz, J.R. 1972. *Joseph Fourier 1768–1830. A survey of his life and work, based on a critical edition of his monograph on the propagation of heat, presented to the Institut de France in 1807*, Cambridge, MA: MIT Press.

Herivel, J. 1975. *Joseph Fourier*, Oxford: Clarendon Press.

Herivel, J. (ed.) 1980. *Joseph Fourier lettres inédites 1808–1816*, Paris: *Bibliothèque Nationale*.

Hewitt, E. and Hewitt, R.E. 1979. 'The Gibbs–Wilbraham phenomenon: an episode in Fourier analysis', *Archive for history of exact sciences, 21*, 129–160.

Hobson, E.W. and Diesselhorst, H. 1904. 'Wärmeleitung', in *Encyklopädie der mathematischen Wissenschaften*, vol. 5, pts. A–B, 161–231 (article V 5).

Macagno, E. 1971. 'Historico-critical review of dimensional analysis', *Journal of the Franklin Institute, 292*, 391–402.

Paplauskas, A.B. 1966. *Trigonometricheski ryadi ot Eulera do Lebega*, Moscow: Nauka.

Weiss, B. 1988. *Zwischen Physiotheologie und Positivismus. Pierre Prevost (1751–1839) und die korpuskularkinetische Physik der Genfer Schule*, Frankfurt: Peter Lang.

CHAPTER 27

JEAN VICTOR PONCELET, *TRAITÉ DES PROPRIÉTÉS PROJECTIVES DES FIGURES*, FIRST EDITION (1822)

Jeremy Gray

Poncelet's book is the source of the launch of the modern study of projective geometry, after the work of various 17th-century authors had not generated sufficient momentum to sustain the subject through the 18th century. He produced a new way of thinking about plane figures that emphasised the properties they have in common with their shadows and played down their metrical properties. He claimed to find many remarkable transformations between figures that enabled complicated figures to be simplified and geometry to work at a new level of generality.

First publication. *Traité des propriétés projectives des figures: ouvrage utile à ceux qui s'occupent de la géométrie descriptive et d'opérations géométriques sur le terrain*, Paris: Bachelier, 1822. xlvi + 427 pages. Print-run: 800 copies.

Second augmented edition. 2 vols., Paris, Gauthier–Villars, 1865–1866.

Related articles: Monge (§17), von Staudt (§33), Riemann on geometry (§34).

1 PONCELET'S *TRAITÉ*

1.1 Principles

Jean Victor Poncelet (1788–1867) discovered many of the key ideas in his book during the two years he was a prisoner of war in Saratov, having been captured by the Russians during the Napoleonic army's retreat from Moscow; as we shall see, this contributed to the book's highly idiosyncratic character. Its impact is all the more remarkable because several of his fundamental ideas and methods were judged by his contemporaries to be unsound, and many of the details of his work had to be replaced entirely.

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

Poncelet arranged for 800 copies of the book to be printed at Metz at his own expense, hoping, as he said towards the end of his life, no other ambition 'but to reach the working class and the youth of our schools' [Poncelet, 1862, vi]. Its contents is summarised in Table 1.

Table 1. Contents by Chapters of Poncelet's book. 'a/b' signifies Section a, Chapter b.

Chap.	**P.**	**Topics**
Preface	i	History of the author's researches.
	iii	Cauchy's *Académie* report on the manuscript, 5 June 1820.
Introd.	xvii–xlvi	Status of geometry: synthetic proofs, principle of continuity. Literature review: mainly Greek and French authors (Pappos, Desargues, Monge).
1	1	'*General principles*'.
1/1	3	Central projection; projective relations.
1/2	26	'Secants and ideal chords' of conic sections; poles and polars. Orthogonality; reciprocal points.
1/3	51	Principles of 'projection of plane figures'. Centres of projection; some special cases.
Notes	71	1) On imaginary limit circles. 2) On a particular projection.
2	76	'*Fundamental properties of straight lines, of conic sections and of circles*'.
2/1	76	Rectilinear geometry and transversals.
2/2	99	In- and escribed figures to conic sections. Reciprocal poles and polars.
2/3	126	'Centre of similitude' in general, and for two circles. Intersecting and touching circles. Similar conic sections.
3	155	'*On systems of conic sections*'.
3/1	156	'Centre of homology'; projection of plane figures, especially conic sections. Applications.
3/2	191	'Complete system' of secants and tangents common to two coplanar conic sections. Systems of them with common secants and tangents.
3/3	228	Double contacts of conic sections; related problems.
4	256	'*On angles and on polygons*'.
4/1	256	Angles of which the corners bear upon the focus, the perimeter of conic sections, or any point of the plane.
4/2	290	On polygons in- and escribed to other polygons or to conic sections.
4/3	329	Theory when directrices are curves of any order, and where certain angles are constant. Applications.
Suppl.	360	Projective properties of figures in space. Homology, continuity.
	417	Table of contents. [End 426.]

In its first chapter Poncelet brought up the key idea: central projection. This is most simply thought of as casting the shadow of a figure in one plane onto another plane by means of a point source of light. The shadow of a straight line is another straight line. The shadow of a circle is some sort of conic section, and with a little work one can see that the shadow can be an ellipse, a parabola, or a hyperbola. Conversely, the shadow of any conic may be a circle, and this suggests that at least some properties of conics can be proved by considering only the circle. But what properties of a conic could these be? It is a familiar experience that a line and its shadow may be of different lengths, and that the shadow of an angle may be an angle of a different size, so shadows do not have the same shape as their originals. The metrical properties of figures are not preserved by projection. The challenge for Poncelet was to find significant properties of figures that are not metrical.

Poncelet found such properties in the incidence of figures. If one line crosses another, then their shadows will cross. If a line touches a curve at some point, then the shadow of the line will touch the shadow of the curve at the corresponding point. If a line cuts a curve in two points, then the shadow of the line will cross the shadow of the curve in two points, from which it follows (with a little work) that the property of being a conic section is a projective property. These were elementary observations, discovered and doubtless forgotten several times in the previous centuries. While still in Saratov, however, Poncelet had found a striking new result, known ever since he published it as Poncelet's porism, and which influenced his decision to continue with his researches. For a valuable modern account of this porism, also known as 'Poncelet's closure theorem', see [Bos et alii, 1987].

Poncelet's porism concerns two conic sections, one inside the other. It says: pick a point A on the outer conic, draw a tangent from it to the inner conic, and continue until it meets the outer conic again, at the point B, say. Repeat the construction, drawing a tangent from B to the inner conic and continuing it until it meets the outer conic again, at the point C say, and carry on doing this. Then either you never return to the point A, or the construction closes up after a finite number of steps, and in this case it will always close up after that many steps no matter what choice you make of the initial point. (The word 'porism' is traditionally attached to construction problems having, as here, an unexpected number of solutions.) We know from his notebooks that Poncelet initially discovered his porism at the end of an immense series of calculations, exactly the sort of long laborious method that he found it impossible to remember; it seems that his realisation that the result might have a much simpler proof when treated by the methods of projective geometry inspired him to develop those methods.

1.2 Projections

With the porism as his motivation, Poncelet proceeded to push the study of the effects of central projection much further than anyone had taken it before. In the preface to the *Traité* he explained that he regretted that the generality of algebra was not matched by a similar generality in geometry. In algebra, and in analytic geometry when geometry is treated algebraically, quantities represented by letters may be added and multiplied without regard to their sign, but in pure geometry it matters whether a point falls inside or outside a segment. Geometric arguments fall into a number of separate cases accordingly, and, as he put it, 'one is forced to reproduce the entire series of primitive arguments from the moment

where a line and a point have passed from the right to the left of one another, etc.' (p. xxii). To avoid this while not surrendering to the algebraists, Poncelet took a radical step. One may savour it in his own words (pp. xxii–xxiii):

> Let us consider an arbitrary figure in a general position and indeterminate in some way, taken from all those that one can consider without breaking the laws, the conditions, the relationships which exist between the diverse parts of the system. Let us suppose, having been given this, that one finds one or more relations or properties, be they metric or descriptive, belong to the figure by drawing on ordinary explicit reasoning, that is to say by the development of an argument that in certain cases is the only one which one regards as rigorous. Is it not evident that if, keeping the same given things, one can vary the primitive figure by insensible degrees by imposing on certain parts of the figure a continuous but otherwise arbitrary movement, is it not evident that the properties and relations found for the first system, remain applicable to successive states of the system, provided always that one has regard for certain particular modifications that may intervene, as when certain quantities vanish or change their sense or sign, etc., modifications which it will always be easy to recognize *a priori* and by infallible rules? [...]
>
> Now this principle, regarded as an axiom by the wisest mathematicians, one can call *the principle* or *law of continuity* for mathematical relationships involving abstract and depicted magnitudes.

One only sees what a radical idea this was when one sees how Poncelet employed it. In the second chapter of the book he took a conic and a chord (a straight line segment defined by a line meeting the conic at two points A and B say). This chord has a midpoint, call it C. Now, varying the figure by insensible degrees, move the line to a new position $A'B'$ parallel to the old one, and locate the midpoint of the new chord, which may be called C'. Continue in this fashion, and an infinite sequence of chords is obtained, together with their midpoints which, it is not difficult to prove, all lie on a straight line that crosses the conic at two points, P and Q, say. Now continue to move the line that cuts out the chord, always keeping it parallel to itself, until it lies entirely outside the conic. Even in this position, said Poncelet, having regard for certain particular modifications, there will still be a chord, this chord will have a midpoint, and this midpoint will also lie on the line PQ. So, according to Poncelet, by means of the law of continuity one may talk of a conic having a chord on a line even when that line does not seem to meet the conic. This eliminates the need for two proofs, one when the line really does meet the conic and one where it does not (as one must say if one does not agree with Poncelet's approach). The chord lying outside the conic he agreed could be called imaginary, the secant ideal; but he insisted that it could be said to meet the conic in ideal points having a real midpoint. Poncelet did show, however, that if the conic is an ellipse there is a unique hyperbola one can draw (satisfying certain requirements glossed over here) with this crucial property: if a line defines an imaginary chord of the ellipse with a midpoint at a specific point, then the line defines a real chord of the hyperbola with its midpoint at the same specific point.

1.3 Continuity

Poncelet had presented exactly this line of reasoning in the memoir of 1820, which he submitted to the *Académie des Sciences* for publication. As was the custom of the time, the memoir was sent out to referees, in this case a panel of three: F. Arago, S.D. Poisson and A.L. Cauchy in the chair. They found much to like in the work, but they had this to say about the law of continuity (as reprinted in the *Traité*, ix):

> This principle, it should be said, is only a bold induction, by means of which one can extend theorems, initially established with certain restrictions, to the case where these restrictions no longer hold. Applied to curves of the second degree, it leads the author to exact results. Nonetheless, we think that it should not be admitted generally and applied indifferently to all sorts of questions in geometry, nor even in analysis.

As author of the report, Cauchy went on to show how everything Poncelet wanted from his law of continuity could be provided by some simple algebra involving complex numbers (compare §25.3). If one chooses coordinate axes, then a conic and a line may be represented by equations. Solving these equations simultaneously, one finds the coordinates of the two points where the line meets the conic. In the case where the line 'really' does, these coordinates are real numbers, and in the case where the line does not 'really' meet the conic (and Poncelet spoke of an ideal line and an imaginary chord) the coordinates are complex numbers. In either case, however, the coordinates of the midpoint are real numbers, and indeed the midpoint so obtained does lie on the line PQ. But this was algebra, and Poncelet wanted geometry. He had no intention of abandoning his project, which he sincerely believed had advantages for students, and instead he kept the method, made it fundamental to the *Traité*, and simply repeated the report in its entirety at the front of the *Traité*, between the Preface and the Introduction—one of the most ostentatious cases in mathematics of deliberately not taking advice.

The first use that Poncelet had for his extended sense in which a line may meet a conic was to the theory of poles and polars, which was a topic already known to geometers, having its roots in Apollonius's study of conic sections. Given a conic C and a point P outside it, draw the two tangents from P to the conic, touching at T and T' say. The line TT' is called the polar line of the point P and the point P is called the pole of the line. If the point P moves along a straight line l it can be shown that the corresponding polar lines all pass through a common point, Q say. This common point Q is called the pole of the line l, which is said to be its polar line. To Poncelet there was no need to distinguish the case where the pole is inside or outside the conic, and he simply proclaimed the theory of poles and polars as an application of his law of continuity.

Now the theory of poles and polars is a striking one, for it allows one to replace a line in a figure by a point simply by introducing an arbitrary conic and proceeding, as above, to replace the point with its polar line. One may also replace a line with the point that is its pole, and it was already known before Poncelet wrote his *Traité* that this process of replacement replaces collinear points with concurrent lines and concurrent lines with collinear points. It is also clear that if the replacement process is conducted twice, it returns the original figure: a point has a polar line and the pole of that line is the original point. The

full import of this idea, called the method of reciprocal polars (or, today, duality) because it puts points and lines in a reciprocal relationship, was to be drawn by many later writers and to become a cornerstone of the theory of projective geometry, but even in Poncelet's *Traité* it is a striking example of a non-metrical geometric property. It is clear from the construction, for example, that the central projection of a conic, a point, and its polar line will be a conic, a point, and a line which is the polar line of that point.

Central projection has a number of unexpected properties. It was a familiar fact from the theory of central focussed perspective in painting that a line in one plane may correspond to a line seemingly infinitely far away in another plane. In this way intersecting lines in one plane may appear as parallel lines in the other plane, and parallel lines may be correctly depicted as ones that intersect. Loose talk of this kind did not strike any mathematician of the 19th century as problematic, but merely as a convenient way of speaking about a simple geometrical fact. Applied to poles and polars, the pole of a line at infinity with respect to a conic is the centre of that conic, and the polar line of the centre of a conic is the line at infinity. This is one of the ways in which talk about a line at infinity enabled geometers to eliminate the need for discussing special cases of certain important results.

But Poncelet did not confine his use of the law of continuity to cases where it was intuitively clear what was going on and where alternative, if more conventional, methods to the same end might seem to lie close at hand. He used in ways that posed successively higher levels of difficulty for any conventional understanding. Not only did a line and a conic always meet, in his enlarged sense of the term, whether or not they seemed to; Poncelet also asserted that a conic and a line can be transformed by a central projection into a circle and a line at infinity. However, sending a conic and a line to a circle and a line at infinity by a projection, or even a sequence of projections, is distinctly problematic. As with the earlier example, it is one thing if the line and the conic do not 'really' meet. Then Poncelet's claim can be proved by conventional means. But if they 'really' meet, the claim is, by conventional standards, false. Once again, Poncelet bought generality by stretching the meaning of the term 'meet'. He did so in order that theorems about a conic and a line (and therefore the theory of reciprocal polars) can be reduced to theorems about a circle and a line at infinity.

Still worse, Poncelet proclaimed that his law of continuity allowed him to treat two conics simultaneously in remarkable ways. The study of the central projection of two conics at once was a new idea of his, and he proclaimed that any two conics may be projected into two circles. This would have struck all his readers as palpably false, for two conics may cross in four points, but two distinct circles may only cross in two points. In the *Traité* Poncelet argued his way round this problem by showing that two of the common points are real, and two are ideal. In the same spirit he proclaimed the special case that two conics tangent to each other at two points are projectively equivalent to a pair of concentric circles. This is equally bizarre, for there seems to be no way in which the points of tangency can be made to disappear. The mathematician who reached for algebra to try to understand what Poncelet was saying would discover, after quite some work, that the only way this can be done is to project the curves not from a real point but from a point with complex coordinates! Such an analysis, which would have raised more problems than it could solve for most of the 19th century, was not available to Poncelet's readers, who were left with only one alternative to following his account, and that was to reject the work entirely.

Readers who persisted would discover that the bulk of the novelty in Poncelet's work consisted of theorems about pairs of objects: either a conic and a line or two conics (as, for example, the porism). There were theorems about conics touching each other, and how to find the central projection transforming one given figure into another. These bore on a classical problem, completely solved by Newton using projective methods almost in passing in his *Principia mathematica* (1687), of how to find a conic given five pieces of information: for example, how to find a conic through five given points, or through three given points and having a given tangent at a given point. The book ends with a variety of construction problems, including the porism, and some remarks about the projective study of quadric surfaces in space.

If one could grant the 'law of continuity' and follow it to the striking conclusions that Poncelet deduced using it, then it permitted a huge simplification of the resulting theory. Configurations could be as simple as possible: a circle and a line at infinity, or, perhaps, two concentric circles. This simplification enabled Poncelet to prove a large number of new results in geometry, and so establish once and for all that there were interesting and valuable non-metrical properties of figures. In short, his achievement was to show that there was a new way of doing mathematics. His methods were not well designed, his fundamental simplifications were not acceptable, but his new subject was.

2 JEAN VICTOR PONCELET AND THE OTHER FRENCH GEOMETERS

Part of the explanation for this extraordinary state of affairs lies in the character of Poncelet himself and the circumstances in which he wrote the book. (For a biography, see [Tribout de Morembert, 1936].) He was a graduate of the *École Polytechnique*, which had been founded in 1794 as part of the educational reforms unleashed by the French revolution. The aim of the *École Polytechnique* was to train would-be engineers for the specialist engineering schools, and after 1804 it became a military school, and many of its graduates were associated with the successes of the Napoleonic army. This gave further status to mathematics within French higher education. While at the *École Polytechnique* from 1807 to 1810, Poncelet came under the influence of Gaspard Monge, one of the founders and the first Director of the school, and of his disciples who were inspired by Monge's vision of geometry. Poncelet graduated from the school in 1810 and moved to the military engineering and artillery school at Metz until 1812; he joined the army as a lieutenant of the engineers just in time to take part in Napoleon's defeat outside Moscow in 1812. He was wounded and left for dead at the battle of Krasnoy, but recovered enough to spend two years as a prisoner of war in Saratov.

To keep up his spirits and those of his fellow prisoners, Poncelet tried to remember what he had learned at the *École Polytechnique;* he published his extensive notebooks in later books [Poncelet, 1862, 1864]. He found he could remember basic results, but not those requiring long laborious methods and what he called 'abstract and spiny proofs'. His fundamental idea of using central projection was a generalisation of a technique emphasised by Monge. Monge had introduced new methods into the study of descriptive geometry, making use of vertical and horizontal projections of figures in three-dimensional space onto the so-called plan and elevation planes. This was fundamental to the use of geometry in architecture and the design of fortifications, and it led in his hands to some simple

useful mathematics, but it was not a profound mathematical discovery (§17). The power of central projection animated Poncelet's search for the simple general methods that led him to write the flawed masterpiece that is his *Traité*, and which he hoped, as he wrote in his [1862], would make geometry 'useful to the working class and the youth of our schools; [and] inspire them with a love of the eternal truths of science'.

On his return from Russia in 1814 Poncelet of course discovered that others had not been idle, and that some of his results were already known. By 1820 he had managed to write up the introductory part of his new system of ideas as a Memoir, in which the 'law of continuity' allowed him to eliminate much of the technical work he had learned to despise while a prisoner-of-war. He knew very well that the algebraic route to these theorems was long and laborious, for his notebooks were full of such proofs. So the response of Cauchy, Arago and Poisson, all men much more influential than he in the mathematical life of the times, angered him but did not make him re-think his programme. He placed it at the front of his *Traité* simply to show that he had no intention of taking their advice.

His bold approach appealed to those who liked geometry, a group that included many former pupils of Monge and others attracted to his synthetic, geometric style of reasoning. But it was less popular with those drawn to the algebraic or analytic branches of mathematics, and many of these people were in positions of influence at the *École Polytechnique*, the specialist engineering schools, or the *Collège de France*. The result was that Poncelet felt his work marginalised and undervalued. He did not deal with this fact of academic life very efficiently, however, and this led him into disputes even with those who might have otherwise admired his work, most notably Joseph Diaz Gergonne (1771–1859).

Gergonne, like Poncelet, was much influenced by Monge, but from 1795 he was a professor in Nîmes and there he set up the first journal devoted exclusively to pure mathematics, the *Annales de mathématiques pures et appliquées*, which ran from 1810 to 1832. During its life it was a major source of articles on projective geometry, and many of the technical terms in the subject appeared there first, often in articles by Gergonne himself ('duality' is one such example). But Gergonne was more sympathetic to algebraic methods in geometry than Poncelet wished, and the two clashed over questions of priority. In the end Poncelet took his articles to other journals, including the German journal *Journal für die reine und angewandte Mathematik*, recently founded by A.L. Crelle in 1826.

The principle of duality had been used by another geometer, C.J. Brianchon (1783–1864), as early as 1806 to deduce the dual of Pascal's theorem. Pascal's theorem says: If A, B, C, D, E, F are six points on a conic, and the lines AB and DE meet at the point P, the lines BC and EF meet at the point Q, and the lines CD and FA meet at the point R, then the points P, Q, and R lie on a line. To state Brianchon's dual result we use the convention that the symbol ab stands for the point common to the lines a and b. Brianchon's theorem says: If a, b, c, d, e, f are six lines touching a conic, and the line joining ab and de is the line p, the line joining bc and ef is the line Q, and the line joining cd and fa is the line r, then the lines p, q, and r meet in a point.

In fact, as the geometers of the time realised, duality allows one to take any projective theorem and not only to state the dual result but also to obtain its proof by dualising the proof of the original result step by step. The dual result might coincide with the original

one, or it might be the converse result, but often it is new, as was the case with Brianchon's dual of Pascal's theorem. Inspired by the fecundity of duality, in 1825 Gergonne proclaimed it as a fundamental principle of plane projective geometry. He argued that it was always possible to take a statement in projective geometry, switch the terms point and line, switch the terms collinear and concurrent, and obtain a new statement. In his view this could always be done, whether or not there was a conic section present. Poncelet disagreed, and the resulting dispute, which was as much about priority in the use of duality as about its inner nature, ran through a number of issues of Gergonne's *Annales* before it petered out. It is interesting to note that mathematically, Poncelet is correct: duality in plane geometry is always duality with respect to a conic. But in the three-dimensional geometry A.F. Möbius later showed that there are dualities (in this case between points and planes in space) that do not require or invoke a conic.

Gergonne then weakened his own position by making a mistake upon which Poncelet pounced. Gergonne considered the process of taking a curve C and a fixed conic E. Each tangent to the curve C is a line, and that line has a pole with respect to the conic E. These poles form a new curve C', which is the dual of the curve C. He stated that if the curve C is defined by a polynomial equation of degree n then so is the dual curve. This is true for conics, which are curves of degree 2; but, as Poncelet saw at once, this is false for curves of higher degree. For example, one may in general draw 6 tangents to a cubic curve from a point not on it (allowing, as one must, for imaginary tangents). The dual of that situation is 6 collinear points on the dual curve, showing that the degree of the dual curve is 6, not 3.

Gergonne was content to admit the mistake, and to dismiss it with the introduction of a new term to handle what was going on; but Poncelet saw that in correcting Gergonne's mistake he had been led to a genuine difficulty. If a cubic curve has a dual of degree 6, a similar though more complicated argument shows that a curve of degree 6 has a dual of degree 30. But it is clear that the dual of the dual of the curve C can only be the original curve C. Since 3 is not equal to 30, there is a paradox to be resolved. Poncelet was not able to resolve it, although he did suspect, correctly, that specific properties of a curve, such as double points and cusps, can lower the degree of the dual. It was left to the German mathematician Julius Plücker to come to the same realisation and prove it, thus opening the way to the projective study of curves of higher degree and the classification of their singular points.

Plücker's marriage of projective and algebraic geometry was a great success. Building as it did on earlier, equally algebraic work by Möbius, it established two things: the power of projective methods to open up new and neglected areas of geometry, and the fact that projective geometry had passed from France to Germany and from synthetic to a mixture of synthetic and more algebraic methods. By this time Poncelet was ready to leave the field he had done so much to create. He had been persuaded by Arago in 1824 to become the professor of mechanics applied to machines at the military school in Metz, and he took to the task energetically (one reason why he published rather tardily on geometry and became embroiled in priority disputes). By the time he moved to Paris in 1837 his interests had shifted completely to experimental mechanics, and projective geometry was left to make its way without him until the 1860s.

3 MICHEL CHASLES

Poncelet had shown conclusively that there was an extensive body of non-metrical results in geometry. But his methods, resting as they did on his 'law of continuity' were by and large not acceptable to his contemporaries. The man who rescued the subject for later generations was Michel Chasles (1793–1880), who had retired in his thirties on a sufficient private income to study mathematics and its history. In 1829 he won a prize from the Belgian Academy of Sciences for a philosophical examination of the different methods of modern geometry, in particular the theory of reciprocal polars. Chasles argued that the theory of projective geometry rests essentially on the notion of cross-ratio, which is preserved by an projective transformation. The cross-ratio of four points A, B, C, D in order on a line can be defined in many ways. Here we take it to be the ratio of ratios $(AB/CB)/(AD/CD) = (AB/BC)/(AD/DC)$, which can be re-written as $(AB \cdot CD)/(AD \cdot CB)$. This was an idea introduced but not stressed by Poncelet, who observed that it could be found in the writings of the Hellenistic geometer Pappus and in modern writers. But Poncelet preferred to rely on the concept of a harmonic division, which is the special case when the cross-ratio equals -1, and the pairs of points A, B, and C, D are related this way: $(AB/BC) = -(AD/DC)$. In this case, the points A and C are said to harmonically separate the points B and D. This is the case that occurs naturally, so to speak, for example in poles and polars and complete quadrilaterals.

Before publishing his prize essay as a book Chasles expanded both the historical introduction and the notes on recent work, and his famous *Aperçu historique* was published as [Chasles, 1837]. It became and remains the first source to consult when thinking about the history of projective geometry, although it is not without its faults (Chasles could not read German). It brought him nothing but success. He was elected to the French *Académie des Sciences* in 1839 as a corresponding (i.e. junior) member. In 1841 he was appointed to the *École Polytechnique*, where he taught geodesy, astronomy, and applied mechanics until 1851, and from 1848 he had a personal chair at the Sorbonne where he taught higher geometry. He was therefore ideally positioned to replace Poncelet's extraordinary methods with his own more rigorous ones, and that is what happened. In the French context, his use of cross-ratio, developed at length in his *Théorie de géométrie supérieure* (1852) with nothing more exotic than imaginary points, did a lot to establish that projective geometry was a legitimate, rigorous discipline in no way dependent on the law of continuity, the theory of ideal points, and other mind-stretching devices used by Poncelet.

In the 1860s Poncelet, now in his seventies, began to feel that his major work on geometry had not been sufficiently appreciated. He re-issued the *Traité* with a few new passages and a second volume of old and new related material [Poncelet, 1862, 1864], and shortly before published a two-volume *Applications d'analyse et de géométrie*, which comprised not only his notes from his time as a prisoner-of-war but a lengthy series of commentaries about all who had, or more often had not, appreciated his work. The most interesting addition to the *Traité* is his recognition that Girard Desargues had had many of these ideas before him. This had been hidden from Poncelet in 1822 by the lamentable publication history of Desargues's masterwork, which had disappeared apparently without trace, leaving only the somewhat scattered comments in the literature to show that it had ever existed.

But in 1845 Chasles had found a hand-written copy made by de la Hire in 1679, and so Poncelet could appreciate what Desargues had done.

The modern historical consensus is undoubtedly that Poncelet brought projective geometry back to life in France. That judgement is fair, even though the methods of projective geometry that succeeded are at least as much Chasles's creation. Chasles also followed tradition by giving the major share of the credit for originality to Monge, with portions also for Brianchon and Gergonne. What that view plays down is that it was Poncelet and only Poncelet who had the vision of a new theory of geometry, as general in its methods as algebra, and which did not need algebra to proceed. This revivified geometry had its own important problems, and a productive method of dealing with them, based on the ingenious use of transformations to reduce the general figure to a simple one. This vision surpassed that of Monge, and if it did not succeed until it was domesticated, a process due in France to Chasles and to others in Germany, it was begun by Poncelet alone. Poncelet's importance to the development of mathematics is secure, but there has been very little historical work done on him, apart from a serviceable biography [Tribout de Morembert, 1936]; and much could be done, because he also had a distinguished career as an engineer.

BIBLIOGRAPHY

Bos, H.J.M., Kers C., Oort F. and Raven, D.W. 1987. 'Poncelet's closure theorem', *Expositiones mathematicae*, 5, 289–364.

Chasles, M. 1837. 'Aperçu historique sur l'origine et le développement des méthodes en géométrie [. . .] suivi d'un Mémoire de géométrie [. . .]', *Mémoires sur les questions proposées par Académie Royale des Sciences et Belles-Lettres de Bruxelles, 11*, 571 pp.

Chasles, M. 1852. *Traité de géométrie supérieure*, Paris: Mallet-Bachelier. [Repr. Paris: Gauthiers–Villars, 1875; Sceaux: Gabay, 1989.]

Kötter, E. 1901. 'Die Entwickelung der synthetischen Geometrie von Monge bis auf Staudt (1847)', *Jahresbericht der Deutschen Mathematiker-Vereinigung, 5*, pt. 2, 486 pp.

Poncelet, J.V. 1862, 1864. *Applications d'analyse et de géométrie* [. . .], 2 vols., Paris: Mallet-Bachelier.

Schönflies, A. 1909. 'Projective Geometrie', in *Encyklopädie der mathematischen Wissenschaften*, pt. 4, sec. 1, Leipzig: Teubner, 389–480 (article IIIAB5).

Tribout de Morembert. H. 1936. *Un grand savant. Le général Jean-Victor Poncelet, 1788–1867*, Paris: Suffroy. [With a portrait.]

CHAPTER 28

A.-L. CAUCHY, TWO MEMOIRS ON COMPLEX-VARIABLE FUNCTION THEORY (1825, 1827)

The late F. Smithies

In these memoirs Cauchy began to lay the foundations of the general theory of analytic functions of a complex variable and their integration by residues.

1825 memoir

First publication. Mémoire sur les intégrales définies, prises entre des limites imaginaires, Paris: de Bure, 1825. 69 pages. Print-run: 500 copies.

Manuscript. Archives of the *Académie des Sciences,* file for meeting of 28 February 1825 (with some deleted passages).

Photoreprint. With Cauchy's textbook on the calculus (§25), Paris: ACL Editions, 1987.

Reprints. 1) *Bulletin des sciences mathématiques, (1) 7* (1874), 265–304; *(1) 8,* 43–55, 148–159. 2) In *Oeuvres complètes,* ser. 2, vol. 15 (1974), 41–89.

German translation. In *Ostwalds Klassiker der exakten Wissenschaften,* no. 112 (ed. and notes P. Stäckel), Leipzig: Engelmann, 1900.

1827 memoir

First publication. 'Mémoire sur les intégrales définies', *Mémoires présentés par divers savants à l'Académie des Sciences, (2) 1* (1827), 601–799. [Mainly written 1814.]

Reprint. In *Oeuvres complètes,* ser. 1, vol. 1 (1882), 319–506.

Related articles: Lagrange on the calculus (§19), Cauchy on real-variable analysis (§25), Riemann on complex-variable analysis (§34).

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

1 BACKGROUND TO THE 1814 MEMOIR

Cauchy's biography was sketched in §25.1, 6, in connection with his founding of real-variable analysis in the 1810s and 1820s; for a full study, see especially [Belhoste, 1991]. On these two memoirs, and their context, see especially [Smithies, 1997].

In discussing these writings of Cauchy and his contemporaries, we shall modernise their notations and terms a little; for example, as a rule we shall say 'complex' rather than 'imaginary', and write 'i' for $\sqrt{-1}$ and '$\frac{\partial f}{\partial x}$' rather than '$\frac{df}{dx}$' for partial derivatives. We shall also use Joseph Fourier's notation '$\int_a^b f(x)\,dx$' for the definite integral, although Cauchy did not deploy it until about 1822.

Towards the end of the 18th century, there was considerable interest in finding ways to evaluate definite integrals, especially when no primitive function was available. Euler [1781] evaluated integrals such as

$$\int_0^\infty x^{m-1}e^{-px}\cos qx\,dx \tag{1}$$

by using substitutions involving complex functions; Laplace [1782] used similar devices to evaluate a family of integrals such as

$$\int_0^\infty \frac{\cos t + t\sin t}{1+t^2}\,dt. \tag{2}$$

From 1809 onwards Laplace used this method of 'imaginary substitutions' several times, mainly for integrals arising in his work on probability theory. This triggered a lively exchange of views with S.D. Poisson (1781–1840), who maintained in a number of papers that the method should be regarded only as a kind of induction, useful for discovering new results, but that these should be confirmed more directly; Laplace defended the method by appealing to the widely accepted principle of the generality of analysis, but eventually had to admit that direct confirmation was desirable.

It seems not unlikely that Cauchy's 1814 memoir was stimulated by a suggestion from Laplace that Cauchy should investigate the method of imaginary substitutions. In his introduction Cauchy refers to the use of the method by Euler and Laplace as a kind of induction from the real to the imaginary, and announces that he proposes to establish it by a direct and rigorous analysis.

2 'CAUCHY'S THEOREM' FORESHADOWED

The memoir is divided into two Parts (Table 1) and is cited by Sections. Cauchy opens Part I by considering a function $f(z)$ of a variable z, which in turn is a function, generally complex-valued, of a pair (x, y) of real variables, and shows that

$$\frac{\partial}{\partial y}\left[f(z)\frac{\partial z}{\partial x}\right] = \frac{\partial}{\partial x}\left[f(z)\frac{\partial z}{\partial y}\right]. \tag{3}$$

(He actually used y as a function of the real variables x and z.) His proof assumes that the functions can be differentiated as often as necessary, even when z takes complex values,

Table 1. Contents of the 1814 memoir.

The pages in the *Oeuvres* printing are given, as it is much more accessible. The first column indicates Parts by roman numerals, Sections therein by the normal ones, and the supplements by ordinals. Cauchy wrote y as a function of real variables x and z.

	Page	**Description**
	329	'Introduction': review of the findings.
I	336	'On the equations which authorise the passage from the real to the imaginary'.
1	336	'General explanation of the method': Cauchy–Riemann equations.
2	339	'First application': integrands include $\exp(-x^{2k})$, x^n and z.
3	349	'Second application': integrands include ax, xz, x^{n-1} and $\exp(-x^2-(m^2/x^2))$.
4	357	'Third application': integrands include $e^x \sin z$ and $e^x \cos z$.
5	369	'Fourth application': ax^2 and xz.
6	362	'On the separation of the exponential' from other functions: examples.
II	378	'On the difficulties which the integration of differential equations can offer'.
1	378	'Double integrals' with an 'indeterminate form' due to infinite values of the integrand.
2	388	Difference in the values taken by the associated repeated integrals in these cases.
3	400	Converting indefinite to definite integrals with infinite integrands.
4	406	'On the value, in finite terms', of the difference in Section 2 above.
5	420	'First application' to Part 1, Section 2 with indeterminacy; examples.
6	463	'Second application' to Part 1, Section 3.
7	465	'Third application' to Part 1, Section 6.
1st	477	'Developments' of Part 2: discussion of two groups of evaluation.
2nd	493	(Supposed) reconciliation of one of his integrations with Legendre's. [End 506.]

and takes for granted the equality of mixed partial derivatives such as $\frac{\partial^2 z}{\partial x \partial y}$ and $\frac{\partial^2 z}{\partial y \partial x}$; behind this lies the usual 18th-century concept of a function as being given by an analytic expression, and thus having a derivative except perhaps at isolated singular values of the variable (compare §19).

We shall not follow Cauchy's discussion in detail; instead we shall restrict our attention to one important special case (Section 2). This will, it is hoped, clarify the structure of his argument. We shall take $z = x + iy$, whereas Cauchy makes the more general assumption that $z = M(x, y) + iN(x, y)$, where M and N are real-valued. In our special case, (3)

reduces to

$$\frac{\partial}{\partial y} f(z) = i \frac{\partial}{\partial x} f(z). \tag{4}$$

If

$$f(z) = P(x, y) + i\,Q(x, y), \tag{5}$$

where P and Q are real-valued, our equation becomes

$$\frac{\partial P}{\partial y} + i \frac{\partial Q}{\partial y} = i \frac{\partial P}{\partial x} - \frac{\partial Q}{\partial x}. \tag{6}$$

Equating real and imaginary parts, we obtain

$$\frac{\partial P}{\partial x} = \frac{\partial Q}{\partial y} \quad \text{and} \quad \frac{\partial P}{\partial y} = -\frac{\partial Q}{\partial x}, \tag{7}$$

the well-known 'Cauchy–Riemann' equations, which actually first appeared in print in an essay by Jean D'Alembert in 1752 (Section 1).

If we now integrate these equations over the domain $x_0 \leqslant x \leqslant X$, $y_0 \leqslant y \leqslant Y$, and note that the double integrals can be evaluated as repeated integrals in either order, we obtain

$$\int_{y_0}^{Y} \big[P(X, y) - P(x_0, y)\big]\, dy = \int_{x_0}^{X} \big[Q(x, Y) - Q(x, y_0)\big]\, dx \tag{8}$$

and

$$\int_{x_0}^{X} \big[P(x, Y) - P(x, y_0)\big]\, dx = -\int_{y_0}^{Y} \big[Q(X, y) - Q(x_0, y)\big]\, dy. \tag{9}$$

It is at the point corresponding to this stage that Cauchy terminates his argument in the general case; in our special case we shall pursue it a little further. If we add i times the first equation to the second one, we shall obtain, with a little rearrangement, the equation

$$\begin{aligned} &\int_{x_0}^{X} \big[P(x, Y) + i\,Q(x, Y) - P(x, y_0) - i\,Q(x, y)\big]\, dx \\ &\quad = \int_{y_0}^{Y} i\big[P(x, y) + i\,Q(X, y) - P(x_0, y) - i\,Q(x_0, y)\big]\, dy, \end{aligned} \tag{10}$$

which we can immediately recognise as saying that $\int f(z)\, dz$, when taken round the sides of the rectangle, is equal to 0. In fact, what we have obtained, is 'Cauchy's theorem' for the special case of a rectangle (Section 2, later footnote). His results in the general case can be interpreted in a similar way as being equivalent to 'Cauchy's theorem' for a kind of general curvilinear quadrilateral.

There are two reasons for Cauchy terminating his argument where he does. Firstly, he is systematically splitting every complex equation into its real and imaginary parts, thus avoiding the use of complex integrands, a procedure that he did not regard as permissible

until about 1819. Secondly, he is carefully avoiding the use of any geometrical language, as he continued to do, following Lagrange's example, until 1825. In spite of these precautions, he was able to use his results to obtain the usual consequences that follow more naturally from the full form of 'Cauchy's theorem'.

The remainder of Part I of the memoir is mainly devoted to applying his results to the evaluation of definite integrals, often over an infinite interval. For instance (Section 2), Cauchy shows that

$$\int_0^\infty e^{-x}\cos ax\,dx = \frac{1}{2}\sqrt{\pi}e^{-a^2/4}. \tag{11}$$

Most of his applications rest on the special case $z = x + iy$ that we examined.

Cauchy's avoidance of complex integrands has the interesting consequence that he cannot use the familiar device of replacing $\cos ax$, where $a > 0$, by e^{iax}, ensuring that $e^{iaz} = e^{ia(x+iy)}$ is small in absolute value when y is large and positive. Instead, in the final Section of Part I, he describes a special device, which he calls 'the separation of exponentials', to get round the difficulty (Section 6).

3 AN ARGUMENT FORESHADOWING THE NOTION OF PRINCIPAL VALUE

An important section of Part II is concerned with the relation between an integral of the form $\int_a^b \phi'(x)\,dx$ and the function $\phi(x)$ in cases where $\phi'(x)$ has an infinity in the interval (a, b). At this time it would generally be assumed that the integral was equal to the difference $\phi(b) - \phi(a)$ between the values of $\phi(x)$ at the ends of the interval; that it was equal in some sense to the sum of the infinitesimals $\phi'(x)\,dx$ was regarded as a theorem that had to be proved. Some paradoxes had arisen in this context over the years; in a typical case, Lagrange [1804, lecture 8] considered the function

$$\phi(x) = \frac{1}{a-x} - \frac{1}{a}, \tag{12}$$

where $a > 0$. Here $\phi(0) = 0$ and $\phi'(x) = 1/(a-x)^2 > 0$; on the other hand, $\phi(x) < 0$ when $x > a$, contradicting the idea that $\phi(x)$ was the sum of the positive infinitesimals $\phi'(x)\,dx$. Lagrange concluded that in this case the principles of the differential calculus were defective.

Cauchy remarks that if $\phi(x)$ varies 'in a continuous manner' in the interval (a, b), we shall indeed have

$$\int_a^b \phi'(x)\,dx = \phi(b) - \phi(a), \tag{13}$$

as we should, his phrasing adumbrates his later definition of continuity for real-valued functions ([Cauchy, 1821]: see §25.4). He suggests that if $\phi(x)$ suffers an abrupt change of value when x passes through X, we should argue as follows; if ξ is small and positive,

we shall have approximately

$$\int_a^b \phi'(x)\,dx = \int_a^{X-\xi} \phi'(x)\,dx + \int_{X+\xi}^b \phi'(x)\,dx$$
$$= \phi(b) - \phi(a) - \Delta \tag{14}$$

where $\Delta = \phi(X+\xi) - \phi(X-\xi)$ is a correction term, which we shall write, more accurately, as

$$\Delta = \lim_{\xi \to 0+} \big[\phi(X+\xi) - \phi(X-\xi)\big]. \tag{15}$$

For example, in the integral $\int_{-2}^{4} \frac{dx}{x}$ the correction term will be

$$\lim_{\xi \to 0+} \big[\ln \xi - \ln(-\xi)\big] = \ln(-1), \tag{16}$$

giving the integral the value $\log 2$ (Section 3). Cauchy's argument here foreshadows his later formal definition [Cauchy, 1822] of the principal value of an integral whose integrand has an infinity.

4 THE FIRST HINTS OF THE RESIDUE THEOREM

We recall that Cauchy's proof of the main theorem of Part I depended on the fact that a double integral can be evaluated as a repeated integral in either order. In this section of Part II he tackles the problem of what happens when both repeated integrals exist but are unequal (Section 2). Cauchy begins with an integral of the form

$$\int_0^1 \int_0^1 \frac{\partial \phi}{\partial y}\,dx\,dy, \tag{17}$$

where $\phi(x, y)$ becomes indeterminate when $(x, y) = (0, 0)$.

When the two repeated integrals are unequal, the difference A between them may be thought of as a correction term, say

$$A = \int_0^1 dy \int_0^1 \frac{\partial \phi}{\partial y}\,dx - \int_0^1 dx \int_0^1 \frac{\partial \phi}{\partial y}\,dy \tag{18}$$

$$= \int_0^1 dy \int_0^1 \frac{\partial \phi}{\partial y}\,dx - \int_0^1 \big[\phi(x,1) - \phi(x,0)\big]\,dx. \tag{19}$$

He then uses an argument like the one he used for single integrals in the last section; this time he cuts out a small rectangle next to the singularity. His conclusion, in modern terms, amounts to saying that

$$A = -\lim_{\varepsilon \to 0+} \lim_{\delta \to 0+} \int_0^{\varepsilon} \phi(\xi, \delta)\,d\xi. \tag{20}$$

Since Cauchy did not then have an efficient notation to indicate the order in which repeated limits are to be taken, his statement of the result is rather clumsier, but clear enough (Section 3).

As an example, let us take

$$\phi(x, y) = \frac{y}{x^2 + y^2} \tag{21}$$

for $0 \leqslant x \leqslant 1, 0 \leqslant y \leqslant 1$. It is easily verified that the correction term is $A = -\pi/4 - \pi/4 = -\pi/2$, and the repeated limit is

$$-\lim_{\varepsilon \to 0+} \lim_{\delta \to 0+} \left[\arctan \frac{\varepsilon}{\delta} \right] = -\frac{\pi}{2}, \tag{22}$$

as it should be (Section 2). A formula for the correction term in the general case where the singularity is at an interior point of the domain of integration is easily found.

We now return to the argument we used earlier to obtain the value of $\int f(z)\,dz$ taken round the sides of a rectangle. If we write, as before,

$$f(z) = P(x, y) + i\,Q(x, y), \tag{23}$$

and try to integrate the Cauchy–Riemann equations over the rectangle, there will be a correction term for each of P and Q (Section 3). That for P will be

$$\begin{aligned} A &= \int_{y_0}^{Y} dy \int_{x_0}^{X} \frac{\partial P}{\partial y} dx - \int_{x_0}^{X} dx \int_{y_0}^{Y} \frac{\partial P}{\partial y} dy \\ &= -\int_{y_0}^{Y} dy \int_{x_0}^{X} \frac{\partial Q}{\partial x} dx - \int_{x_0}^{X} dx \int_{y_0}^{Y} \frac{\partial P}{\partial y} dy \\ &= -\int_{y_0}^{Y} \big[Q(X, y) - Q(x_0, y) \big] dy - \int_{x_0}^{X} \big[P(x, Y) - P(x, y_0) \big] dx. \end{aligned} \tag{24}$$

Similarly, the correction term for Q will be

$$B = \int_{y_0}^{Y} \big[P(X, y) - P(x_0, y) \big] dy - \int_{x_0}^{X} \big[Q(x, Y) - Q(x, y_0) \big] dx. \tag{25}$$

Combining the two equations, we see that

$$\begin{aligned} A + i\,B &= \int_{y_0}^{Y} \big[f(X + iy) - f(x_0 + iy) \big] i\,dy - \int_{x_0}^{X} \big[f(x + iY) - f(x + iy_0) \big] dx \\ &= \int f(z)\,dz \end{aligned} \tag{26}$$

taken round the sides of the rectangle (Section 5). We have thus found an expression for this integral involving the correction terms for P and Q, i.e. essentially involving the correction term for $f(z) = P + i\,Q$.

Cauchy's next move is to investigate the effect of the simplest possible singularity; he supposes that $f(z)$ has a simple pole at $z = \alpha + i\beta$, say, and no other singularities. He expresses $f(z)$ in the form $F(z)/G(z)$, where $F(z)$ is well behaved and $G(z)$ has a simple zero at $\alpha + i\beta$, and writes

$$F(\alpha + i\beta)/G'(\alpha + i\beta) = \lambda + i\mu. \tag{27}$$

If $z = (\alpha + \xi) + i(\beta + \eta)$, where ξ and η are small, then

$$f(z) = \frac{\lambda + i\mu}{\xi + i\eta} + u(z), \tag{28}$$

where $u(z)$ has no singularity; the correction term for $f(z)$ will therefore be the same as that for

$$\phi(\xi + i\eta) = \frac{\lambda + i\mu}{\xi + i\eta} \tag{29}$$

at $\xi + i\eta = 0$ (Section 4). Using earlier results, we get the answer $2\pi i(\lambda + i\mu)$, which we can recognise at once what Cauchy was later to call the residue of $f(z)$ at the pole $z = \alpha + i\beta$.

As before, we have restricted ourselves to the special case $z = x + iy$, whereas Cauchy works through the details of the general case $z = M(x, y) + iN(x, y)$. This involves him in some very complicated calculations for the correction terms. Nevertheless, he finds that when $f(z)$ has a simple pole, the correction term for $f(z) = F(z)/G(z)$ is still equal to $2\pi i F(\alpha + i\beta)/G'(\alpha + i\beta)$, exactly as before; thus, to his expressed surprise, it is independent of the functions M and N (Section 4, after equation (22)). This seems to indicate that he has not yet fully understood the situation.

The remaining portions of the memoir are mainly concerned with the application of Cauchy's results to the evaluation of definite integrals. In particular, he evaluates a large family of principal-value integrals, a typical one being

$$\int_0^\infty \frac{\sin ax}{\sin bx} \frac{dx}{1 + x^2} = \frac{\pi}{e} \frac{e^a - e^{-a}}{e^b - e^{-b}}, \tag{30}$$

where $0 < a < b$. In his report as one of the referees of the memoir, Legendre asked why the condition $a < b$ was required; Cauchy emphasised in his reply that it was necessary for his proof that

$$\frac{\sin a(x + iy)}{\sin b(x + iy)} \cdot \frac{1}{1 \cdot (x + iy)^2} \tag{31}$$

should be small in absolute value for large x and y (Suppl. 2).

We can sum up the achievement of the 1814 memoir briefly by saying that in it Cauchy has proved a primitive form of 'Cauchy's theorem' and taken the first steps towards the theory of residues.

5 THE NOTES ADDED BEFORE PUBLICATION

Before the 1814 memoir was ultimately published in 1827, Cauchy adjoined a group of footnotes to it; these seem to date from about 1825. Their main thrust is that he abandons his earlier rule only to consider integrals with real integrands; we have already seen that many of his results can be stated more simply if complex integrands are admitted. He remarks that throughout the memoir many pairs of real equations can be replaced by a single complex equation. There are also some simplifications in the proofs of his results. For instance, in considering integrals of the form $\int_0^\infty x^{n-1}F(x)\,dx$ in Part I, he originally worked in terms of the real and imaginary parts of $F(x)$, which led him to some very complicated expressions; he now works directly with the expression $\int_0^\infty (x-ib)^{n-1}F(x)\,dx$ (Section 2, equation (11)). Again, he now found it possible to dispense with the special device of 'separation of exponentials' that he had introduced to cope with integrals of the form $\int f(x)\cos ax\,dx$, and to work directly with $\int f(x)e^{iax}\,dx$ (Section 6, equation (32)).

6 THE BACKGROUND OF THE 1825 MEMOIR

At the beginning of our discussion of Part II of the 1814 memoir we mentioned the paradoxes that occasionally arose in identifying $\int_a^b \phi'(x)\,dx$ with $\phi(b)-\phi(a)$ when $\phi'(x)$ has an infinity in the interval (a,b). Poisson [1820] is concerned with some of these. He suggests that the validity of the equation

$$\int_{-1}^{1}\frac{dx}{x} = (2n+1)\pi i \tag{32}$$

could be re-established by making the change of variable

$$x = -(\cos\rho + i\sin\rho) \tag{33}$$

and replacing the limits -1 and 1 by 0 and $(2n+1)\pi$; here he has encountered a situation where, as we should put it, the value of a definite integral depends on the path along which it is evaluated. He treats the integral $\int_{-1}^{1} dx/x^m$ in a similar way for a positive integer m.

In another example Poisson considers the integral

$$y = \int_{-\infty}^{\infty}\frac{\cos ax}{b^2+x^2}\,dx \tag{34}$$

where $a>0$ and $b>0$, and examines the effect of the substitution $x = t+ki$, where k is a real constant. He finds that

$$y = \frac{\pi}{b}e^{-ab}, \quad \text{where } 0<k<b \tag{35}$$

but

$$y = \frac{\pi}{2b}\left(e^{-ab}-e^{ab}\right), \quad \text{where } k>b. \tag{36}$$

Table 2. Contents of the 1825 memoir. 68 pages.

Art(s)	Description
1	Introduction: singular integrals.
2–3	Concept of complex-variable function '$f(x + y\sqrt{-1})$' and its integral.
4–8	Case when function takes infinite values: evaluation of its integral, residues.
9–11	Curves in the complex plane specified by values of (x, y); straight line.
12	Residues: evaluation of integrals, many particular cases.
13	Infinitude of residues; evaluations.
14–15	Further evaluations, including summation of infinite series.
16	Curve defined in contiguous sectors specified by different functions.
17–18	Multiple complex integrals; example from hydrodynamics.

In effect, the path of integration has been moved from the real axis to a parallel straight line. Cauchy mentions [Poisson, 1820] in his paper [Cauchy, 1822] on principal-value integrals, and it looks very much as if this paper of Poisson's was the stimulus that provoked Cauchy into writing the 1825 memoir (Table 2).

7 DEFINITE INTEGRALS BETWEEN COMPLEX LIMITS

Cauchy begins the 1825 memoir with a first attempt at a direct generalisation of his 1823 definition of the definite integral of a function of a real variable (§25.4). He says that, with $z = x + iy$, then $\int_{x_0+iy}^{X+iY} f(z)\,dz$ should be the limit (or one of the limits) of sums of the form

$$\begin{aligned}&\big[(x_1 - x_0) + i(y_1 - y_0)\big]f(x_0 + iy_0) + \big[(x_2 - x_1) + i(y_2 - y_1)\big]f(x_1 + iy_1)\\ &\quad + \cdots + \big[(X - x_n) + i(Y - y_n)\big]f(x_n + iy_n),\end{aligned} \tag{37}$$

where the sequences $(x_0, x_1, \ldots, x_n, X)$ and $(y_0, y_1, \ldots, y_n, Y)$ are monotonic (either increasing or decreasing) and the differences $(x_k - x_{k-1})$ and $(y_k - y_{k-1})$ tend to zero as n increases indefinitely (art. 2). The monotonicity requirement is obviously copied from the real-variable case.

Cauchy seems to realise at once that this definition is not specific enough; hence the phrase 'one of the limits'. To remedy this defect, he suggests the following way of constructing such sequences: let $x = \phi(t)$, $y = \psi(t)$, where $\phi(t)$ and $\psi(t)$ are continuous monotonic functions for $t_0 \leqslant t \leqslant T$, $\phi(t_0) = x_0$, $\psi(t_0) = y_0$, $\phi(T) = X$, $\psi(T) = Y$, and write $x_k = \phi(t_k)$, $y_k = \psi(t_k)$, where $(t_0, t_1, \ldots, t_n, T)$ is a monotonic increasing sequence. The sum introduced above now becomes an approximating sum for the integral

$$\int_{t_0}^{T} \big[\phi'(t) + i\psi'(t)\big] f\big[\phi(t) + i\psi(t)\big]\,dt, \tag{38}$$

which we take to be $\int f(z)\,dz$ along the path defined by $x + iy = \phi(t) + i\psi(t)$ for $t_0 \leqslant t \leqslant T$.

This change in the way he defines the integral is typical of the whole memoir; throughout it, Cauchy seems to be changing his mind as he proceeds. This perhaps gives one some insight into the working of his mind.

8 THE MAIN THEOREM OF THE 1825 MEMOIR

In the next section of the memoir Cauchy assumes that $f(x + iy)$ is finite and continuous for x between x_0 and X and y between y_0 and Y. He takes it for granted that $f'(x + iy)$ exists and is continuous, His announced aim is to prove that the value of the integral $\int_{x_0+iy_0}^{X+iY} f(z)\,dz$ is independent of the choice of the functions $x = \phi(t)$ and $y = \psi(t)$; in other words, that the integral will have the same value for all monotonic paths from $x_0 + iy_0$ to $X + iY$. There is an analogy here with his recent proof that the integral of a real-variable function is independent of the sequence of partitions used in defining it (§25.4).

Essentially Cauchy gives three proofs for his result. The second proof is couched in the language of the calculus of variations; supposing that the functions $x = \phi(t)$ and $y = \psi(t)$ are given small variations δx and δy, he obtains

$$\delta \int_{t_0}^{T} (x' + iy')\,f(x + iy)\,dt = \int_{t_0}^{T} \left[(x' + iy')\delta f(x + iy) + f(x + iy)\delta(x' + iy')\right] dt = 0 \tag{39}$$

(art. 3). In other words, a small variation in the path of integration has a zero affect.

His first proof spells out the second one in much more detail. His third argument is a remark that the result could have been foreseen, since $f(x + iy)(dx + i\,dy)$ is an exact differential; that is, if we write

$$f(x + iy)(dx + i\,dy) = U\,dx + V\,dy, \tag{40}$$

then

$$\frac{\partial U}{\partial y} = \frac{\partial V}{\partial x} = if'(x + iy). \tag{41}$$

Whether he intended this remark as an alternative proof is not clear. In more modern terms, his argument seems to involve the continuous deformation of the path of integration, suggesting a (not quite rigorous) homotopy argument; it was made more rigorous in a brief note by M. Falk in 1853.

9 TAKING SINGULARITIES INTO ACCOUNT

Cauchy next attacks the problem of what happens when $f(z)$ has a singularity at a point lying between two paths along which the integral $\int_{x_0+iy_0}^{X+iY} f(z)\,dz$ is evaluated, so that the conditions for the two values of the integral to be equal are no longer satisfied (arts. 4–5). Here again he gives several alternative treatments, appearing to change his mind while preparing the memoir. At this stage he adopts some suggestions made to him by a young

Russian mathematician, M.V. Ostrogradsky (1801–1862), who was working in Paris at the time; his assistance is acknowledged by Cauchy in his introduction to the memoir.

Cauchy deals first with the case where the singularity is a simple pole, giving it two separate treatments. In the first of these he uses an approximation argument, getting involved with some complicated, expressions, which he ploughs through in his usual manner. In the second treatment he splits $f(z)$ into two parts, one if which is a simple rational expression (now sometimes called the principal part of the function), reproducing the behaviour of $f(z)$ near the singularity, and the other a well-behaved function, which makes no contribution to the difference between the integrals. We omit the details, only remarking that the device of splitting the function in this way appears to come from Ostrogradsky, who had used a similar device in an unpublished paper dated 24 July 1824 and was working closely with Cauchy at the time [Yushkevich, 1965].

Cauchy then goes on to deal with the case where the pole of $f(z)$ between the paths is a multiple one (arts. 6–7). Again he gives two treatments; in the first one he again uses Ostrogradsky's device of splitting a function into a rational principal part and a well-behaved function, and in the second he gives an alternative approximation argument of ferocious complexity. Again we omit details, remarking only that the inclusion of this approximation is odd, since the use of the principal part enables him to evaluate some auxiliary expressions explicitly; it looks as if the second proof for a simple pole and the first proof for a multiple pole were inserted at a late stage in the preparation of the memoir. In every case he concludes that the difference between the integrals is $\pm 2\pi i f_0$, where f_0 is precisely what he was later to call the 'residue' of $f(z)$ at the singularity.

10 THE USE OF GEOMETRICAL LANGUAGE

Shortly after proving these results Cauchy suddenly introduces, without any prior warning except for a brief mention in his preliminary abstract, some geometrical language in describing his results (art. 9). In this essay we have used it freely in our discussion of both the 1814 and the 1825 memoirs, but up to this point Cauchy had carefully avoided it; as with Lagrange, he had mistrusted its use, feeling that it would tend to disguise the general validity of analytic methods. From this point onward, he began to use it, occasionally to start with, and in later papers more and more freely. He seems to feel, indeed, that some of the ideas appearing in his present context can be expressed more concisely by admitting some geometrical terminology; it enables him, for instance, to say that a singularity lies 'between' two paths.

He approaches geometrical ideas by remarking that, if one eliminates t from the equations $x = \phi(t)$, $y = \psi(t)$, one will obtain an equation of the form $F(x, y) = 0$; if (x, y) are regarded as rectangular coordinates in the plane, this equation will represent a curve joining the points (x_0, y_0) and (X, Y). He also goes on to say that the complete path may be made up of several portions defined by different functions, provided only that each separate portion satisfies his conditions (art. 16). We also remark that he does not use the full force of J.R. Argand's geometrical representation [1806] of complex numbers; he never mentions the geometrical interpretations for addition and multiplication.

The remainder of the memoir need not concern us in detail; he gives numerous illustrative examples of the evaluation of particular definite integrals and the derivation of identi-

ties between them. He also examines some cases where the function $f(z)$ has an infinite number of singularities, using them for the summation of certain infinite series.

We see that in this memoir Cauchy has achieved primitive versions of all the basic results of complex function theory, but there was still some way to go for these results to be put in a digestible form.

11 LATER DEVELOPMENTS

It is a well known phenomenon in the history of mathematics that, when the initial proofs of major theorems are at all complicated, it will be the task of the pioneer's successors to improve and simplify the results and develop their ramifications. In Cauchy's case he was in great part his own successor. In 1826 he began to publish a series of papers in his own journal *Exercices de mathématiques*, in which he defined the residue of a function at a singularity and built up a calculus of residues. At first he treated the residue theorem for rectangles as fundamental; from 1827 onwards he shifted the emphasis from rectangles to circles, which enabled him to improve his treatment of the behaviour of a function $f(z)$ for large $|z|$ and eventually, in the first of two memoirs presented to the Turin Academy in 1831, to give simple proofs of convergence for the Maclaurin and Taylor series of analytic functions; he also discussed power series representing implicit functions, such as the Lagrange series. In his second Turin memoir, Cauchy proved the residue theorem for a simple closed curve: he also introduced what he called the 'index' of a function with respect to a contour, a notion closely related to what was later to be called its winding number.

After 1831 Cauchy used his complex analysis to obtain existence theorems for differential equations and for other applications. However, he contributed nothing essential to the general theory until he resumed giving regular lectures after the 1848 revolution.

One naturally asks how these memoirs of Cauchy and their immediate sequels were received. At first they aroused very little interest, for the exciting growth of the theory of elliptic and related functions was found much more attractive than Cauchy's ideas. Apart from a few minor notes by G. Piola, B. Tortolini and one or two others on the residue calculus, the first contributions to the general theory were P.A. Laurent's 1843 paper in the *Comptes rendus* of the Paris Academy on the power series expansion for a function with an isolated essential singularity, and Liouville's 1844 theorem, also in the *Comptes rendus*, that a bounded analytic function is necessarily a constant; as Cauchy promptly pointed out, both results are easy consequences of his work. Victor Puiseux's important memoir in Liouville's *Journal de mathématiques* on algebraic functions came out in 1850.

We saw earlier that in both the 1814 and the 1825 memoirs, and indeed for a long time afterwards, Cauchy took it for granted that a continuous function, even of a complex variable, has a derivative, which in general will also be continuous. It was not until Cauchy resumed regular lecturing, probably about 1851, that he realised that an expression of the form $P(x, y) + iQ(x, y)$, where $P(x, y)$ and $Q(x, y)$ are arbitrary continuous functions of the pair of real variables (x, y), has to be regarded as a continuous function of $z = x + iy$, and that for it to be differentiable as a function of z, it must satisfy the Cauchy–Riemann equations. His detailed conclusions can be found in [Cauchy, 1853]. Riemann made the same point in his 1851 Göttingen dissertation (§34).

The atmosphere changed substantially with the appearance of Riemann's dissertation and the growing influence of Weierstrass's Berlin lectures after 1857, together they brought about a tremendous development of complex function theory. In particular, Cauchy's work on complex analysis began to be understood and appreciated at its true value; the first connected account of Cauchy's ideas appeared in a memoir by C. Briot and J. Bouquet in the *Journal de l'École Polytechnique* in 1856, followed by their book on doubly periodic functions of 1859 and a book by F. Casorati in 1868. From then onwards Cauchy's work on complex function theory was generally accepted as a fundamental contribution to the structure of the subject.

BIBLIOGRAPHY

Argand, J.R. 1806. *Essai sur une manière de représenter les imaginaires dans les constructions géométriques*, Paris: the author.

Belhoste, B. 1991. *Augustin-Louis Cauchy: a biography*, New York: Springer.

Bottazzini, U. 1990. 'Editor's introduction', to A.-L. Cauchy, reprint of *Cours d'analyse de l'École Royale Polytechnique*, Bologna: Cooperativa Libraria Universitaria Editrici, ix–clxvii.

Cauchy, A.-L. 1821. *Cours d'analyse de l'École Royale Polytechnique*, Paris: de Bure [Repr. in *Oeuvres complètes*, ser. 2, vol. 3. See also [Bottazzini, 1990], and §25.]

Cauchy, A.-L. 1822. 'Memoire sur les intégrales définies', *Bulletin de la Société Philomatique de Paris*, 161–174. [Repr. in *Oeuvres*, ser. 2, vol. 2, 283–299.]

Cauchy, A.-L. 1853. 'Sur les différentielles des quantités algébriques ou géométriques et sur les dérivées de ces quantités', *Exercices d'analyse et de physique mathématique*, vol. 4, 336–347. [Repr. in *Oeuvres*, ser. 2, vol. 4, 393–406.]

Euler, L. 1781. 'De valoribus integralium a termino variabilis $x = 0$ usque ad $x = \infty$ extensorum', in *Institutiones calculi integralis*, vol. 4, 337–345. [Repr. in *Opera omnia*, ser. 1, vol. 19, 217–227. German trans. in *Ostwalds Klassiker der exakten Wissenschaften*, no. 261 (1983), 229–239, with notes and commentary by A.P. Yushkevich.]

Lagrange, J.L. 1808. 'Leçons sur le calcul des fonctions', *Journal de l'École Polytechnique, 5, cahier 12*. [Repr. as *Oeuvres*, vol. 10.]

Laplace, P.S. 1782. 'Mémoire sur les approximations des formules qui sont fonctions de très grands nombres', *Mémoires de l'Académie Royale des Sciences*, 1–88. [Repr. in *Oeuvres*, vol. 10, 209–291.]

Poisson, S.D. 1820. 'Suite du memoire sur les intégrales définies', *Journal de l'École Polytechnique, 11, cahier 18*, 293–341.

Smithies, F. 1997. *Cauchy and the creation of complex function theory*, Cambridge: Cambridge University Press.

Timchenko, I. 1899. *Osnovanii teorii analiticheskikh funktsii* ['Foundations of a theory of analytic functions'], Odessa: Shul'tse.

Yushkevich, A.P. 1965. 'O neopublikovannikh rannikh rabotakh M.V. Ostrogradskogo [On unpublished early works of M.V. Ostrogradsky]', *Istoriko-matematicheskie issledovaniya, 16*, 11–48.

CHAPTER 29

NIELS HENRIK ABEL, PAPER ON THE IRRESOLVABILITY OF THE QUINTIC EQUATION (1826)

Roger Cooke

This paper represents one of several early attempts to prove the nonexistence of an algorithm for solving an equation of fifth or higher degree by algebraic operations alone.

First publication. 'Beweis der Unmöglichkeit, algebraischer Gleichungen von höheren Graden als dem vierten allgemein aufzulösen', *Journal für die reine und angewandte Mathematik, 1* (1826), 65–84.

French translation. In *Œuvres complètes*, Christiania: Grøndahl, 1839, vol. 1, 5–24. Also in *Œuvres complètes*, 2nd ed., Christiania: Grøndahl, 1881, vol. 1, 66–87.

English translation. In Peter Pesic, *Abel's proof: an essay on the sources and meaning of mathematical unsolvability*: Cambridge, MA and London: MIT Press, 2003, 155–180.

Related articles: Cauchy on analysis (§25, §28).

1 INTRODUCTION: EQUATIONS IN GENERAL

1.1 Linear equations and negative numbers

Although many of the problem-solving methods we now know as algebra are very ancient, the explicit statement of the notion of an equation, that is, the equivalence of two formally different expressions for the same unknown quantity, was first formulated by Diophantus of Alexandria (dates uncertain) in the second or third century CE. As Bashmakova and Smirnova note [1997, 132], 'Diophantus was the first to deduce that it was possible to formulate the conditions of a problem as equations or systems of equations; as a matter of fact, before Diophantus, there were no equations at all, either determinate or indeterminate. Problems were studied that we can now reduce to equations, but nothing more than that'.

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

Diophantus's work was known to the Muslim mathematicians of medieval times, who also knew of the work of the Hindu mathematicians of the sixth and seventh centuries, and equations are classified and studied explicitly in the work of the Muslim mathematician al-Khwarizmi in the ninth century. The distance from all this early algebra to what is known today as elementary algebra is considerable. The chief difference is the absence of symbolism, and particularly parameters, in the early years to systematize the discussion. Equations were stated in words. It is now possible, using parameters a and b, to state the generic linear equation in one unknown as $ax + b = 0$ and its solution as $x = -b/a$. Although this statement could have been easily explained to al-Khwarizmi, there are two reasons why he could not have made it. First, as already mentioned, he did not have the concept of parameters; second, he did not use stand-alone negative numbers, only subtracted numbers. As a result, he was forced to explain himself by using many examples. It is of particular interest that linear equations in full generality cannot be solved without the use of negative numbers. Mathematicians are therefore forced to make sense of such numbers, or else admit that not every linear equation has a root.

1.2 Quadratic equations: irrational, imaginary, and complex numbers

Solving quadratic equations requires the extraction of square roots, and these lead to other objects not recognized as numbers in the Greek tradition. Al-Khwarizmi, who considered only what we call linear and quadratic equations, classified them into six types: 'squares plus numbers equal roots', 'squares plus roots equal numbers', 'squares equal roots plus numbers', and three other forms in which a term is missing, the latter including the case of what we call linear equations. Since some square roots could be known only approximately as numbers, al-Khwarizmi gave geometric solutions of quadratic equations. Here again, one sees that equations force the consideration of certain kinds of numbers (irrational square roots) that the Greek mathematicians would have referred to as *magnitudes* rather than *numbers*.

The representation of magnitudes as line segments was arithmetized to an extent by René Descartes (1596–1650), who in his *Géométrie* showed how to represent the product of two line segments as a line segment by choosing a fixed segment to represent unity; he also gave a geometric representation for the square root of a number (§1). In all but name, Descartes's approach made it possible to think of irrational magnitudes as numbers; and a generation later Isaac Newton (1642–1727) defined a number to be the ratio of one magnitude to another magnitude of the same kind, arbitrarily taken as a unit. Newton classified numbers as integers, fractions, and surds [Whiteside, 1967, 7]. Worse things than irrationalities arise in the case of quadratic equations, however, since the procedure for finding the roots sometimes involves the square root of a negative number. Since the roots of such equations are not what we now call real numbers, it was possible to ignore these cases without forgoing the solution of any equation that was regarded as solvable.

1.3 Cubic and quartic equations

Later Muslim mathematicians, such as Omar Khayyam (1056–1130), similarly classified equations containing the third power of the unknown and showed how a solution of such

an equation could be represented as the intersection of two conic sections, but a formula for the solution involving only rational operations and root extractions was not found until the 16th century.

This discovery produced two striking effects. First, it was soon followed by a formula for solving the general quartic equation. Second, it turned out that the solution of an equation with real coefficients always required the extraction of the square root of a negative number when there are three real roots. The formula given by Girolamo Cardano, Niccolò Tartaglia, and others, for example, when applied to the equation $x^3 + 6 = 7x$, gives the solution

$$x = \sqrt[3]{3 + \sqrt{\frac{-100}{27}}} + \sqrt[3]{3 - \sqrt{\frac{-100}{27}}}. \tag{1}$$

Mathematicians could and did call such solutions *impossible*. However, that impossibility left unsolved an equation that definitely did have solutions (namely $x = 1$, $x = 2$, $x = -3$). Because there were real solutions in this case, known as the 'irreducible' case, an effort was made to find them.

François Viète (1540–1603), who was also the first to state explicitly the fact that the coefficients of a polynomial are symmetric functions of its roots, succeeded in solving this case using only real numbers by appeal to the trigonometric formula

$$4\cos^3\theta - 3\cos\theta = \cos(3\theta). \tag{2}$$

Viète's solution, however, was not algebraic, since it involved transcendental functions. The problem of what to do with square roots of negative numbers was investigated by Rafael Bombelli (1526–1573) and others in the 16th century, and some progress was made by John Wallis (1616–1703) and others in the 17th century in providing a geometric interpretation for what we now call complex numbers. As in the cases of negative and irrational numbers, the problem of which numbers one can sensibly talk about—to use a mathematical colloquialism, the problem of which mathematical objects *exist*—arose from the need to make sense of an algebraic formula. There is a natural inclination to make sense of any operation that arises naturally in a formula. That natural impulse has led to the acceptance of negative, irrational, and imaginary quantities as numbers capable of being added, subtracted, multiplied, and divided, and obeying (most of) the usual laws of arithmetic.

One bit of 'transcendentalism' has always remained in the irreducible case, however: While the extraction of the square root of a complex number can be reduced to the extraction of the square roots of positive real numbers, the situation is different for cube roots. It is not possible to reduce the extraction of the cube root of a complex number to algebraic operations involving only real numbers. The equations for the real and imaginary parts of the cube root of a complex number are themselves cubic equations having in general three real roots, and hence belong to the 'irreducible' case. They can, however, be expressed using cube roots of real numbers and trigonometry. More generally, the nth root of a complex number can be expressed as the nth root of its absolute value times a complex number of unit length whose polar angle is $1/n$ times the polar angle of the given number. As Ayoub [1980, 254–255] notes, mathematicians have been content to take the existence of roots in the complex numbers as given.

1.4 The 'fundamental theorem of algebra'

Because of the algebraic/trigonometric representation of roots of complex numbers, it was clear that any algebraic formula requiring only rational operations and root extractions would not lead to any new numbers beyond the complex numbers. It came to be generally believed that any polynomial equation with complex roots would have a solution in the complex numbers (the property now known as algebraic closure of the complex numbers). The first explicit statement of this theorem occurs in the book *L'invention nouvelle en l'algèbre* (1629, Amsterdam), written by Albert Girard (1595–1632). Girard came to the conclusion that an equation of degree n with real coefficients has n roots in the complex numbers. One can imagine that the conviction that complex roots exist for all equations was reinforced by the tacit assumption that a *formula* involving only rational operations and root extractions could be found which, when applied to the coefficients, would produce a solution.

1.5 A general technique for solving equations

The method of solving quadratic equations is known as completing the square. That is, given the equation $x^2 - ax + b = 0$, one makes the substitution, $x = y + a/2$, to get the equation $y^2 = a^2/4 - b$, which is solvable by extraction of the square root. For the general cubic equation $x^3 - ax^2 + bx - c = 0$, the substitution $x = y + a/3$, similarly produces a cubic equation in y in which the square term is missing, and that transformation formed a crucial part of Cardano's solution of the cubic. This generalization of completing the square, however, does not 'complete the cube', since the linear term, in general, remains. The quest for a way to complete the cube and any higher power was begun by Ehrenfried Walther von Tschirnhaus (1652–1708), who thought at first that he had succeeded. In 1677 he wrote to Leibniz [Leibniz, 1850, 429]:

> In Paris I received some letters from Mr. Oldenburg, but from lack of time have not yet been able to write back that I have found a new way of determining the irrational roots of all equations [...]. The entire problem reduces to the following: we must be able to remove all the middle terms from any equation. When that is done, and as a result only a single power and a single known quantity remain, one need only extract the root.

Tschirnaus claimed that the $n - 1$ middle terms of the equation

$$x^n - a_1 x^{n-1} + \cdots + (-1)^{n-1} a_{n-1} x + (-1)^n a_n = 0 \tag{3}$$

could be eliminated by a substitution of the form

$$y = x^{n-1} + b_1 x^{n-2} + \cdots + b_{n-2} x + b_{n-1}, \tag{4}$$

when the coefficients $b_1, \ldots, b_{n-1}$ are suitably chosen. Such a transformation is now called a *Tschirnhaus* transformation. It provides the basis for a proposed method of solving any equation, by induction on the degree. Assuming one can solve all equations of degree $n - 1$ or less, one can express x in terms of y with coefficients that are algebraic functions of

the coefficients $b_1, \ldots, b_{n-1}$. When this expression is substituted into (3) and the radicals cleared, the result is an equation in y with coefficients that are polynomials in the b's. Then by setting all the middle coefficients equal to zero and solving for the b's, one is to obtain an equation of the form $y^n = C$. After y is found by extracting the nth root, the original equation will have been solved, since x is already expressed in terms of y. Tschirnhaus illustrated with the case of the cubic equation $x^3 - qx - r = 0$, taking $y = x^2 - ax - b$. He noted that y satisfied the equation

$$y^3 + (3b - 2q)y^2 + (3b^2 + 3ar - 4qb + q^2 - a^2q)y + \left(b^2 - 2qb^2 + 3bar + q^2b - aqr - a^2qb + a^3r - r^2\right) = 0. \tag{5}$$

Thus one could eliminate the square term and the linear term by first choosing $b = 2q/3$, and then solving for a in the quadratic equation

$$qa^2 - 3ra + q^2/3 = 0. \tag{6}$$

At the very least, Tschirnhaus had found a second solution of the general cubic equation, independent of those given earlier by the Italian mathematicians of the preceding century, and his method certainly seemed to be general. The devil was in the details, however, and those details involved horrendously complicated computations when applied to higher-degree equations.

It appears that Tschirnhaus obtained (3) by solving for x in terms of y, substituting, and then removing the radical by squaring. There is, however, a more general way of obtaining it. If the roots of the original cubic equation are, say, u, v, and w, and f is any polynomial, the coefficients in the equation

$$\big[y - f(u)\big]\big[y - f(v)\big]\big[y - f(w)\big] = 0 \tag{7}$$

are symmetric polynomials in $f(u)$, $f(v)$, and $f(w)$, and hence also symmetric in u, v, and w. They can therefore be expressed as polynomials in the coefficients of the original equation and the coefficients of the polynomial f—as was shown conclusively by Edward Waring (1734–1789) in 1762—so that it is not necessary to know u, v, and w in advance in order to write down (7). When f is the quadratic used by Tschirnhaus, (7) is precisely (5). When this procedure is applied in general, as Tschirnhaus proposed, the equations obtained by setting the middle coefficients equal to zero are of degrees $1, 2, \ldots, n-1$ in the coefficients of f, and can be solved for those coefficients, just as Tschirnhaus had claimed. As an algorithm, however, the process turns out to involve an infinite loop. The coefficients in all but the first of the equations to be solved are *mixed*, that is, expressions of the form $b_i b_j$ occur. When such an equation is solved for one of these two variables, the other occurs as part of the expression for the first *under a radical sign*. Hence the next equation to be solved contains, in addition to the variable not solved for previously, a fractional power of a polynomial containing that variable. When that radical is cleared out, the resulting equation is, in general, of *higher* degree than the original equation. Hence the proposed algorithm requires an even stronger algorithm in order to function. This effect appears even in the case of the quartic equation. However, since the quartic can be reduced to a quadratic by

removing only two of its coefficients, it is still possible to use a Tschirnhaus transformation to solve it.

Despite its ultimate failure, Tschirnhaus's method revealed a great deal about the process of finding solutions of equations. If indeed there is a method of solving every polynomial equation by algebraic operations, then all the steps in Tschirnhaus's procedure could be carried out, even though they would not constitute an inductive procedure for solution. One feels intuitively that if there were an algorithm for systematically solving all equations, it ought to have been the one proposed by Tschirnhaus. Thus, the suggestion that there are equations that cannot be solved algebraically arises from Tschirnhaus's work. Whether or not mathematicians of the time had the same feeling, it was clear that more data, more particular examples, were needed in order to distill a general approach to the solution of equations. But a gain had been made: the work of Tschirnhaus, and Girard had focused attention on the importance of the symmetric polynomials in the roots. In the light of these advances, mathematicians again turned their attention to the problem of solving the quintic equation.

2 THE PROBLEM OF THE QUINTIC, 1700–1800

2.1 Finding a solution

It was Viète who took the important step of using parameters to discuss equations, thereby providing an essential element in the statement of the general problem. After Viète, only a minimal change in notation is needed to write the generic quintic equation as

$$x^5 - ax^4 + bx^3 - cx^2 + dx - e = 0. \tag{8}$$

The problem faced during the 18th century was to find a sequence of rational operations and root extractions, which, applied to the coefficients, would produce a solution. To illustrate the idea, consider the general quadratic equation

$$x^2 - ax + b = 0. \tag{9}$$

The general formula that solves this equation can be reduced to the following sequence of operations:

$$t_1 = a^2 - 4b, \quad t_2 = \sqrt{t_1} = \sqrt{a^2 - 4b}, \quad t_3 = \frac{a + t_2}{2} = \frac{a + \sqrt{a^2 - 4b}}{2}. \tag{10}$$

Here each operation is either a rational operation on the preceding results and the coefficients, or a root of a preceding result. The solution of the general cubic and quartic equation can similarly be reduced to such a sequence of rational operations and root extractions, albeit a much longer and more complicated sequence. Although no one actually wrote out the proposal in these explicit terms, the challenge to the 18th-century mathematicians was to produce such a sequence for the equation (8).

2.2 Euler

In the paper [1738], and again in [1762], Leonhard Euler (1707–1783) tried his hand at solving equations of arbitrary degree. The first time, generalizing from the case of equations of degrees 2, 3, and 4, he asserted that the solution of an equation of degree n could be expressed in the form

$$x = \sqrt[n]{A_1} + \sqrt[n]{A_2} + \cdots + \sqrt[n]{A_{n-1}}, \tag{11}$$

where $A_1, \ldots, A_{n-1}$ are the roots of an equation of degree $n-1$, called by Euler the *resolvent*. The program is that of Tschirnhaus, and the difficulty is the same: Finding the coefficients of the resolvent is not always easier than the original problem. In his second attempt, Euler assumed a solution of the form

$$x = w + A\sqrt[n]{v} + B\sqrt[n]{v^2} + \cdots + Q\sqrt[n]{v^{n-1}}, \tag{12}$$

where w is a real number and v and the coefficients $A, \ldots, Q$ are to be found by forming the analog of (6) and comparing with the original equation. This approach, and a similar approach of Étienne Bézout (1730–1783), also does not lead to any solution of the general quintic equation.

In between these two efforts Euler [1749] considered the existence of roots, showing that an equation whose degree is a power of 2 can be split into two equations of equal degree and incidentally stating the conjecture that the roots of an equation of degree higher than 4 cannot be found with a finite number of algebraic operations.

2.3 Lagrange

A decade after Euler's second attempt, Joseph-Louis Lagrange (1736–1813) wrote a long survey [1772–1773] of the methods known up to his time for solving general equations. Regarding Tschirnhaus's general method, he noted the following (*Œuvres*, vol. 3, 305):

> To apply, for example, Tschirnhaus's method to the equation of degree 5, one must solve four equations in four unknowns, the first being of degree one, the second of degree 2, and so on. Thus the final equation resulting from the elimination of three of these unknowns will in general be of degree 24. But, apart from the immense amount of labor needed to derive this equation, it is clear that after finding it, one will be hardly better off than before, unless one can reduce it to an equation of degree less than 5; and if such a reduction is possible, it can only be by dint of further labor, even more extensive than before.

Lagrange devoted a large amount of space to analysis of the cases when the resolvent equation can be reduced below the degree of the equation it is derived from, based on the ingenious trick of expressing a root of the resolvent in terms of the roots of the original equation and simply counting how many different values the root of the resolvent must assume (in general) when the roots of the original equation are permuted among themselves. He showed, for example, that the resolvent for a general quartic equation, which is

formally of degree six, can be written as the product of three quadratic polynomials whose coefficients have the form $(MP + NQ)/(P^2 + Q^2)$ and $(M^2 + N^2)/(P^2 + Q^2)$, where M, N, P, and Q are expressions of degree 4 in the roots of the quartic assuming only three different values when the roots are permuted. Therefore the resolvent will reduce to a cubic equation. In this idea, he laid the egg that later hatched out as Galois theory. However, a considerable amount of pecking by Ruffini, Cauchy, Abel, and Galois, was needed before the shell broke, as we shall see.

2.4 Gauss

Carl Friedrich Gauss (1777–1855) began his mathematical career with his 1799 dissertation, in which he proved the fundamental theorem of algebra. He preceded his proof with a thorough examination of earlier attempts to prove this theorem and mentioned incidentally Euler's 1749 conjecture on the impossibility of solving the general quintic by algebraic operations. He commented (*Werke*, vol. 3, 17),

> Against this the argument may be advanced that after the efforts of so many mathematicians, there remains very little hope of finding the general solution of arbitrary algebraic equations, so that it seems more and more likely that such a solution is absolutely impossible and contradictory. This may seem less paradoxical if we note that *what is ordinarily called a solution of an equation is properly speaking merely a reduction of the equation to pure equations*. But here the solution of pure equations is not proved, only assumed [...]. It may not be difficult to prove the impossibility with complete rigor for the quintic, and I shall expound my own research on this matter at more length elsewhere.

2.5 Ruffini

The first claim of a proof that it is impossible to find a formula for solving all quintic equations by algebraic operations is due to the Italian scholar Paolo Ruffini (1765–1822), published in the same year, 1799, when Gauss wrote his dissertation. Ruffini's education covered a number of areas, including both mathematics and medicine. In his argument Ruffini made use of Lagrange's counting of the number of values a function can assume when its variables are permuted. Unfortunately, he assumed that the radicals that arise in the course of solving the equation are rational functions of the roots, and this assumption requires proof [Bryce, 1986]. He was unsuccessful in getting the proof recognized by the French Academy, the acknowledged center of mathematical life at the time, even though he revised the proof twice to make it clearer.

2.6 Cauchy

One French mathematician who did recognize Ruffini's achievement and regarded his proof as valid was Augustin-Louis Cauchy (1789–1857). In 1812 Cauchy read an essay on symmetric functions before the Academy. This paper was later published as [Cauchy, 1815]. He proved the important fact that a function of n variables that assumes fewer values

than the largest prime number less than n when the roots are permuted, actually assumes at most two values. This result was a key fact used by Abel in his proof.

3 NIELS HENRIK ABEL

The author of this attempt to demonstrate the impossibility of solving the quintic equation algebraically was born in 1802, into a family of very limited means. At the time Norway, as a dependency of Denmark, was suffering economically from the British efforts to eliminate or neutralize the Danish Navy to keep it out of the hands of Napoleon. In addition, his father was engaged in political activity, agitating for Norwegian independence, and seems to have been a heavy drinker as well, contributing to the family's poverty. In 1815, the year after control of Norway shifted to Sweden, Abel and his older brother were sent to the Cathedral School in Christiania (Oslo). As most of the good teachers from the Cathedral School had gone to provide the teaching staff at the University of Christiania in 1813, Abel found few good teachers there and was not inspired. However, in 1817, a new mathematics teacher, Bernt Holmboe, arrived, and was to prove a constant friend and mentor to Abel. When Abel's father died in 1820, Holmboe helped the young student to obtain a scholarship to continue his education.

The following year Abel entered the University of Christiania. That same year, thinking he had succeeded in solving the quintic equation, he sent a paper to the Danish mathematician Ferdinand Degen, who asked him for a specific example of his method. Working out such an example revealed the mistake to Abel, and a few years later, he wrote the first draft of an impossibility proof, published privately in 1824 by Grøndahl (*Œuvres*, vol. 1, 28–34). In this proof Abel recognized the importance of filling in the gap in Ruffini's work. Unfortunately, his proof that the intermediate radicals in a supposed solution by formula can be expressed as rational functions of the roots suffers from some vagueness also, and the version that he finally published in the *Journal für die reine und angewandte Mathematik* was greatly expanded, with fuller explanations of the use of permutations.

Degen had advised the talented young man to devote himself to elliptic functions, and this area, and its generalization to integrals of completely general algebraic functions formed the vast majority of Abel's life work and perhaps the most profound theorem of the early 19th century, called *Abel's theorem* at the suggestion of his rival in elliptic functions C.G.J. Jacobi (1804–1851) (§31). In 1825, after two years of intensive study of the German and French languages, Abel went on a tour of Denmark, Germany, and France, carrying some of his papers by way of introduction. The trip was only a partial success; he met Crelle, whose *Journal für die reine und angewandte Mathematik* provided the outlet for most of his work. He was less successful in his attempts to meet Gauss, A.M. Legendre, and others, and the referees from the French Academy (Cauchy and Legendre) took little interest in the brilliant paper containing Abel's theorem. (This paper was published only in 1841, after considerable urging by Jacobi.) Abel returned home in 1827, in debt. He died in 1829 at the age of 26, his best work still unrecognized [Stubhaug, 2000]. Ironically, the Paris Academy awarded him a prize for his work on elliptic functions the following year.

4 CONTENTS OF THE MEMOIR

In the introduction Abel states the problem: *To express the roots of the equation as algebraic functions of the coefficients.* Nowadays, this definition would appear circular, since the modern definition of an algebraic function y of variables $x_1, \ldots, x_n$ is simply a function that satisfies an equation $p(x_1, \ldots, x_n; y) = 0$, where p is a polynomial in its arguments. What Abel meant, however, was shown by his definitions in the first section of the paper: it should be possible to write y as the end of a chain of operations, each of which is either a rational operation or the extraction of a root of prime order. (Roots of composite order can be obtained by a succession of extractions of roots of prime order.)

In the first section Abel classifies algebraic operations according to the levels of nesting of roots that it contains. A function whose construction requires only the extraction of roots of rational functions of the variables is of order one. One whose construction requires the extraction of roots of expressions of order $\mu - 1$ or lower is of order μ. An algebraic function is precisely a function having some finite order, and Abel then showed that a general algebraic function v of order μ could always be written in the form

$$v = q_0 + p^{1/n} + q_2 p^{2/n} + q_3 p^{3/n} + \cdots + q_{n-1} p^{(n-1)/n}, \tag{13}$$

where p is an algebraic function of order $\mu - 1$, and n is a prime number. The apparently missing coefficient q_1 in (13) was subsumed into the function p. Abel justified this transformation by saying that one could always replace p by $q_1^n p$, giving a separate argument in the case when $q_1 = 0$. William Rowan Hamilton (1805–1865) was later to point out that doing so could increase the order of this term, since q_1 might itself be of order μ.

The brief second section is devoted to proving that when the expression v in (13) is assumed to be a root of an algebraic equation and is substituted for the unknown in the equation, the resulting algebraic expression, when reduced to the form (13), has all coefficients 0. The argument that proves this fact is extremely clever, and amounts to considering the simultaneous zeros of the polynomials $z^n - p$ and the polynomial that results from the original substitution. (These polynomials have at least the common zero $p^{1/n}$.) At the end of this section, Abel concludes that the solution *can be* given in a form in which all the algebraic functions involved are rational functions of the roots. The phrase *can be* is to be emphasized here. Not every representation of a root will have this property.

The longer third section is devoted to a detailed proof of the result of Cauchy quoted above. In the brochure he had printed in 1824, Abel had merely given a reference to Cauchy's work; here he reproduces the relevant parts of Cauchy's paper, with credit to Cauchy.

The very brief final section wraps up the proof by showing that the hypothetical solution leads to an equation in which one side has 120 different values when the roots of the quintic are permuted, while the other side has only 5 values. This contradiction establishes the theorem.

5 RECEPTION OF THE MEMOIR

The underlying basis of this work—the theory of groups of permutations—was still not fully worked out. Lacking this elegant context, Abel's proof, like Ruffini's, suffers from

a certain vagueness as to what is necessary or possible in a hypothetical solution. As a result, Abel's proof did not gain universal recognition. Both proofs, reclothed in modern notation, can be found in [Ayoub, 1980], who also points out that it is possible *a priori* that every equation of degree five might be solvable by radicals, without there being any single formula for doing so. The impossibility proofs of Ruffini and Abel were directed at the stronger hypothesis of a single formula. Of course, it is now possible to exhibit particular quintics that cannot be solved by radicals, so that even the weaker hypothesis is refuted. Évariste Galois (1811–1832), who knew of Abel's work, clarified the matter, making the important distinction between normal and non-normal subgroups, or, as he called them, proper and improper decompositions of the group (into cosets). In 1832 the Prague Scientific Society declared the proofs of both Ruffini and Abel unsatisfactory and offered a prize for a correct proof [Ayoub, 1980, 274].

In 1839 Hamilton published what the American number theorist and historian of mathematics Leonard Eugene Dickson called 'a very complicated reconstruction of Abel's proof', discouragingly burdened with primed subscripts and superscripts [Hamilton, 1839]. As mentioned above, Hamilton noted that it was crucial to Abel's proof that the coefficient q_1 in (13) is 1. This was a gap in Abel's proof, although Ayoub [1980, 272] notes that it can be repaired; and indeed, Hamilton did repair it. After surveying cubic and quartic equations, Hamilton turned to the quintic, saying that 'the opinions of mathematicians appear to be not yet entirely agreed respecting the possibility or impossibility of expressing a root as a function of the coefficients by any finite combination of radicals and rational functions'.

In Hamilton's proof, as in Abel's, the crucial fact is that all the successive functions obtained in a hypothetical solution must be expressible as rational functions of the roots, admitting a definite number of values as the roots are permuted. Hamilton devoted many pages to writing out all the possible forms of rational functions of three and four variables having specified numbers of values when their arguments are permuted. For the quintic, he showed that there are only two basic types of rational functions having fewer than 6 values when the arguments are permuted, one being the product of the differences of two variables (two values), the other a polynomial of degree 4 in one of the variables (five values).

The ideas of both Abel and Galois were developed further by Laurent Wantzel (1814–1848) and Enrico Betti (1823–1892). Wantzel, who in 1837 had shown the impossibility of doubling the cube or trisecting the angle using ruler and compass, showed in [1845] that it is impossible to solve all equations in radicals. Nowadays all of these impossibility results are derived from group theory. In 1852 Betti published a number of theorems elucidating the theory of solvability by radicals. An analytic (not algebraic) solution of the quintic was published in 1844 by Ferdinand Eisenstein [Patterson, 1990].

Abel's proof came just after the basic tool for answering the question—the symmetric group—had been introduced. Independently of the correctness of his proof by modern standards, his skillful employment of that tool showed its importance and pointed the way toward a full understanding of what is involved in the solution of equations by radicals.

BIBLIOGRAPHY

Ayoub, R. 1980. 'Paolo Ruffini's contributions to the quintic', *Archive for history of exact sciences*, *23*, 253–277.

Bashmakova, I.G. and Smirnova, G.S. 1997. 'The origin and development of algebra', in B.V. Gnedenko (ed.), *Essays on the history of mathematics*, Moscow: Moscow University Press, pp. 94–246. [In Russian. English trans.: *The beginnings and evolution of algebra* (trans. A. Shenitzer), [Washington]: Mathematical Association of America, 2000.]

Betti, E. 1852. 'Sulla risoluzione delle equazioni algebriche', *Annali di matematica pura ed applicata, 3*, 49–51.

Bryce, R.A. 1986. 'Paolo Ruffini and the quintic equation', *Symposia mathematica, 27*, 169–185.

Cauchy, A.L. 1815. 'Mémoire sur les functions qui ne peuvent acquérir que deux valeurs…', *Journal de l'École Polytechnique, (1) 10, cah. 17*, 29–112. [Repr. in *Œuvres complètes*, ser. 2, vol. 1, 91–169.]

Dahan, A. 1980. 'Les travaux de Cauchy sur les substitutions. Étude de son approche du concept de groupe', *Archive for history of exact sciences, 23*, 279–319.

Euler, L. 1738. 'De forme radicum aequationum cuiusque ordinis coniectatio', *Commentarii Academiae Petropolitanae*, 6 (1732), 216–231. [Repr. in *Opera omnia*, ser. 1, vol. 6, 1–19.]

Euler, L. 1749. 'Recherches sur les racines imaginaires des équations', *Histoire de l'Académie des Sciences de Berlin*, 5, 222–288. [Repr. in *Opera omnia*, ser. 1, vol. 6, 78–147.]

Euler, L. 1762. 'De resolutione aequationum cuiusque gradus', *Novi Commentarii Academiae Petropolitanae, 9*, 70–98. [Repr. in *Opera omnia*, ser. 1, vol. 6, 170–196.]

Gauss, C.F. 1799. *Demonstratio nova theorematis omnem functionem algebraicam rationalem integralam unius variabilis in factores reales primi vel secundi gradus resolvi posse*, Helmstedt. [Repr. in *Werke*, vol. 3, 1–30. Part in L. Euler, *Opera omnia*, ser. 1, vol. 6, 151–169.]

Hamilton, W.R. 1839. 'On the argument of Abel, respecting the impossibility of expressing a root of any general equation above the fourth degree, by any finite combination of radicals and rational functions', *Transactions of the Royal Irish Academy, 18*, 171–259. [Repr. in *Mathematical papers*, vol. 3, 517–569.]

Lagrange, J.L. 1772–1773. 'Réflexions sur la résolution algébrique des équations' and 'Suite', *Nouveaux Mémoires de l'Académie Royale des Sciences et Belles-Lettres de Berlin*, (1770), 134–215; (1771), 138–253. [Repr. in *Œuvres*, vol. 3, 205–421.]

Lagrange, J.L. 1795. 'Leçons élémentaires sur les mathématiques données à l'*Ecole Normale*', *Séances de l'Ecole Normale, passim*. [Repr. in *Œuvres*, vol. 7, 181–288. English trans.: *Lectures on elementary mathematics* (trans. T.J. McCormack), Chicago: Open Court, 1898.]

Leibniz, G.W. 1850. *Mathematische Schriften* (ed. C.I. Gerhardt), vol. 4, Halle (Saale): Schmidt. [Repr. Hildesheim and New York; Olms, 1971.]

Ore, O. 1957. *Niels Henrik Abel, mathematician extraordinary*, Minneapolis: University of Minnesota Press. [Repr. New York: Chelsea, 1974.]

Patterson, S.J. 1990. 'Eisenstein and the quintic equation', *Historia mathematica, 17*, 132–140.

Pesic, P. 2003. *Abel's proof: an essay on the sources and meaning of mathematical unsolvability*, Cambridge, MA and London: MIT Press.

Stubhaug, A. 2000. *Niels Henrik Abel and his times: called too soon by flames afar*, Berlin–New York: Springer. [Trans. by Richard Daly of *Et foranskutt lyn: Niels Henrik Abel og hans tid*, Oslo: Aschehoug, 1996.]

Van der Waerden, B.L. 1985. *A history of algebra from al-Khwarizmi to Emmy Noether*, Berlin: Springer-Verlag.

Wantzel, L. 1845. 'Démonstration de l'impossibilité de résoudre toutes les équations algébriques avec des radicaux', *Bulletin des Sciences, par la Société Philomathique de Paris*, 5–7.

Waring, E. 1762. *Miscellanea analytica*, Cambridge: Bentham.

Whiteside, D.T. (ed.) 1967. *The mathematical works of Isaac Newton*, vol. 2, New York and London: Johnson.

CHAPTER 30

GEORGE GREEN, *AN ESSAY ON THE MATHEMATICAL ANALYSIS OF ELECTRICITY AND MAGNETISM* (1828)

I. Grattan-Guinness

Although published very obscurely, this book became recognised as a major text in potential theory after gaining publicity from the 1850s. The integral theorem and function which came to carry Green's name gained especial attention.

First publication. Nottingham: 'printed for the author', 1828. ix + 72 pages.

Photoreprints. Berlin: Mayer und Müller, 1890. Goteborg: Ekelöf, 1958. Nottingham: The University, 1993.

Reprints. In *Journal für die reine und angewandte Mathematik*, *39* (1850), 75–89, *44* (1852), 356–374, *47* (1854), 161–211. [Repr. in *Mathematical papers* (ed. N.M. Ferrers), London: Macmillan, 1871, 1–115. Edition repr. Paris: Hermann, 1903; New York: Chelsea, 1970.]

German translation. Ein Versuch, die mathematische Analysis auf die Theorieen der Elektrizität und des Magnetismus anzuwenden (trans. A.J. von Öttingen and A. Wangerin), Leipzig: Engelsmann, 1895 (*Ostwald's Klassiker der exakten Wissenschaften*, no. 61).

Related articles: Laplace on celestial mechanics (§18), Fourier (§26), Thomson and Tait (§40), Maxwell (§44).

1 GREEN'S LITTLE-KNOWN ENTRÉE

One of the major publications in the mathematics of the 19th century was also one of the most obscure. In Nottingham, a once pleasant small town then succumbing to the effects of the Industrial Revolution, the book appeared by private subscription, with 52 named supporters [H. Green, 1946, 549–552]; the print-run is unknown. While the town was growing

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

commercially, it was no intellectual centre; so the book made no impact until its posthumous (re)discovery, when its importance was quickly recognised. Disclosure of the life of its author has occurred much more recently, especially with the biography [Cannell, 1993]; but no *Nachlass* has been retrieved, and no likeness is known. So his life is exceptionally obscure.

However, Green's personal roots are clear, as the family was an example of the commercial transformation of the period. He was born sometime in July 1793 to a baker after whom he was named; he was to be first of four children. Several years later his father established a corn mill at Sneinton, then a village not far from the town, and for much of his life Green worked in it; indeed, he inherited it after his father's death in 1829, soon after the publication of his book.

No precise information exists on Green's education or the development of his mathematical ability. The Nottingham Subscription Library contained a respectable stock of scientific publications, including even some foreign material. There were men of scientific training in the town, most notably John Toplis, headmaster of a school, who had translated the first Book of Laplace's *Traité de mécanique céleste* (1799) (§18) in 1814 and was well aware of British inferiority in mathematical research. Green did not study at his school but may well have had contact with him. At all events, the references and surrounding remarks in his book showed that he had not only become aware of Laplace and several other French authors but had also read several of their works. In particular, one of Laplace's leading followers had produced a result that may well have *inspired* Green to his research in the first place [Grattan-Guinness, 1995].

2 POISSON'S 'SIMPLIFYING' THEOREM

The domination of mathematics by the French from the 1780s until the time of Green's book is quite extraordinary; most work of importance was coming out of Paris. Moreover, the atmosphere there was professional to a degree previously unknown; new and old institutions of higher education, and chances of employment as a scientist on a scale unavailable elsewhere [Grattan-Guinness, 1990]. One of the new generation to profit from these circumstances was Siméon-Dénis Poisson (1781–1840), associated with the *Ecole Polytechnique* (founded in 1794) all his adult life and active in other educational institutions; and unceasing in his research efforts, most of which focused upon the main topic of the time, the calculus and its applications to mathematical physics (including mechanics). By and large he was conservative throughout his career in that he took as his mentors Laplace and J.L. Lagrange (1736–1813).

One of Poisson's main innovations occurred in the mid 1820s when he attempted the novelty of mathematicising magnetic bodies [Poisson, 1826]. He treated a magnetic body M as composed of separate dipoles and analysed the strength of its attraction to internal and to external monopoles. In his first paper he found a theorem which converted triple integrals over the volume of M to double integrals over its surface, which he regarded merely as a 'simplification' of certain formulae. An English translation of a summary version of the paper appeared in the *Quarterly journal of science* as [Poisson, 1824], and Green may well have seen it: while no mathematical details were provided, it could have excited his

curiosity. When the full paper appeared, he must have been inspired by the theorem, novel at the time and indeed not fully understood by its author. Green realised that this result was not just simplification: far more profoundly, *it related properties inside a body to properties on its surface and vice versa*.

3 GREEN'S THEOREM IN GREEN'S BOOK

Whether or not this result was his source, Green prepared his book on the mathematicisation of electricity and magnetism. The preface is dated March 1828; Poisson's paper was published at the end of 1826, and could have been in Britain by, say, the spring. If indeed Poisson was his inspiration, Green could have written the book in around a year—a fine effort, and not impossible to imagine given the undoubted flow of rich ideas. At 81 pages, the book was quite short; for example, about the length of Poisson's first paper. After a preface and some pages of 'Introductory observations', it divided into three Parts: 'General preliminary results' and then applications to electricity and to magnetism (see Table 1). The page numbers refer to the edition in his collected works, as it is the most accessible version.

Early in the first Part Green produced a theorem of the same kind as Poisson's, but with proper understanding of its physical significance. For two 'continuous functions' $U(x, y, z)$ and $V(x, y, z)$

$$\int dx\,dy\,dz\,U\delta V + \int d\sigma\,U\left(\frac{dV}{dw}\right) = \int dx\,dy\,dz\,V\delta U + \int d\sigma\,V\left(\frac{dU}{dw}\right), \qquad (1)$$

where 'δ' was the Laplacian operator (an unusual choice of symbol, perhaps made by his printer), '$d\sigma$' an element of the surface, all integrals were stated with only one integral sign '$\int$', and round brackets indicated partial 'differential co-efficients' (a practice brought in by Euler). The form of the theorem (1) became known as 'symmetric'. As with Poisson, Green's proof was effected by integrating by parts, in his case on

$$\int dx\,dy\,dz\left\{\left(\frac{dV}{dx}\right)\left(\frac{dU}{dx}\right) + \left(\frac{dV}{dy}\right)\left(\frac{dU}{dy}\right) + \left(\frac{dV}{dz}\right)\left(\frac{dU}{dz}\right)\right\}, \qquad (2)$$

to produce each side of (1) (pp. 24–26). But he did not follow Poisson in modifying the proof to allow for non-convex surfaces by taking such integrals over convex parts as appropriate.

When 'singularities' in U (or V) occurred at points G, examination showed that terms of the form $-4\pi U(x_G, y_G, z_G)$ were needed on the appropriate side of the equation (p. 27); Green also inserted a term of this kind when the attracted point was internal to the body in question (p. 40). These modifications suggest that he was aware of Poisson's modification of Laplace's equation for interior points of attraction; dating from 1813, Poisson had used it again in a succeeding paper on magnetisation [Poisson, 1827], which Green mentioned on p. 6.

These theorems, especially (1), helped Green specify 'the potential function', as he called it (p. 9), and now named after him, namely:

Table 1. Summary of Green's book. The page numbers pertain to the edition in [Green, 1871]. Those to the left start off the Parts. LE = expansion of the potential function in a series of Legendre functions.

Pages	Topics and methods
3	'Preface'. Survey of results in mathematics and physics: Cavendish, Poisson, Laplace.
9	'Introductory observations'. 'Potential function', 'Laplace's equation'; summary of results to come.
19	'General preliminary results'. Laplace's and Poisson's equations; Green's theorem.
27	Surface and interior potentials; Green's function, its existence and uniqueness.
42	Applications to electricity.
42	'Leyden phial': equation for its geometry; several linked phials.
50	Two spheres: LE; 'electric density'; case of one external point.
55	Spheres joined by a 'fine wire'; density across a thin spherical shell.
57	Density on thin spherical shell containing small circular orifice.
61	'[L]ong metallic wires' joining spheres in the atmosphere.
63	Density for conducting bodies in electrical equilibrium; LE.
68	Potential for a straight 'line'.
69	Electricity in imperfectly conducting body in magnetic field.
75	Ditto for electrical field; equipotential surfaces.
83	Applications to magnetism.
83	Potential function for 'very small body'; LE.
87	Magnetic equilibrium for 'any body'; potential function as volume and surface integrals.
95	Magnetism of hollow spherical shell; LE.
100	Potential of an infinite planar plate of uniform thickness; LE.
106	Magnetic distribution in a wire; theoretical calculation and Coulomb's data. [End 115.]

> It only remains therefore to find a function V' which satisfies the partial differential equation, becomes equal to [a given function] $\bar{V}'$ when [the point] p is upon the surface A, vanishes when p is at an infinite distance from A, and is besides such, that none of its differential co-efficients shall be infinite, when the point p is exterior to A.

(p. 12: note that the prime does not denote differentiation and $V'(\infty)$ is not clearly specified.) He claimed that the function was unique (pp. 31–32), although he assumed without argument that it was proportional to $1/r$ for large distance r (for example, p. 33); the uniqueness question was to inspire much perplexed analysis later. He also gave a symmetrical form for the relationship between two such functions (pp. 37–39), launching what has become known as 'reciprocity relations'.

A key result followed: the relationship between density '(ρ)' at a point of the surface and the potential gradient there (p. 32):

$$0 = 4\pi(\rho) + \left(\frac{dU}{dw}\right). \tag{3}$$

The form of this equation recalls Fourier's handling of surface heat diffusion (§26, (7)); perhaps he knew the *Théorie analytique de la chaleur* (1822).

Works in (mathematical) physics which Green explicitly mentioned included Henry Cavendish's studies of the early 1770s (pp. 3–4, but not used later), and Poisson's papers of the 1810s on electricity as well as the later ones on magnetism (pp. 4, 6). In his introduction he compared Fourier with Poisson and A.L. Cauchy on methods of solving differential equations in hydrodynamics (pp. 7–8), which suggests that he had read the comparison made in [Fourier, 1818]. On mathematical methods, maybe the specification of continuous functions in his theorem reflects awareness of the reforms of the calculus then being effected by Cauchy (§25). A passing reference to Lagrange's follower L.F.A. Arbogast shows his familiarity with French techniques using differential operators (p. 103; compare p. 77); he may have learnt of them from the second edition (1810s) of S.F. Lacroix's large treatise on the calculus (§20), for he cited it by page number for another result (p. 113).

The means of Green's access to these—indeed, any—sources is unknown. Perhaps English-language journals read in the subscription library or elsewhere had given him the references, although mathematical literature was less well covered than that of the other sciences. Presumably he had sufficient funds to buy some books and papers: the international market had long been well organised, and suffered interruption only during the last years of Napoléon in the 1810s.

4 GREEN'S APPLICATIONS TO ELECTRICITY AND MAGNETISM

In the rest of the book Green gave a variety of applications and examples of his theorems in electrical or magnetic situations; they are curiously little-studied. Among the 'preliminary results' (3) and its kin feature more than (1). Table 1 lists the topics and methods covered; a few features will be brought out here.

Assuming that potential functions always existed (pp. 78, 92), Green favoured treating them as given by the solution of 'Laplace's equation' (his pioneering use of this name) for points external to the body in question. He developed them in a series of Legendre functions (to use the modern name); Book 3 of Laplace's *Traité de mécanique céleste* (1799) (§18.5) was cited on several occasions as his source. He often truncated this and other power-series to the first or second term when the increment variable was small (for example, pp. 43, 112–113). Some of his cases extended or varied upon those treated by Poisson; like his predecessor, he seems not to have carried out any experiments himself. He also did not discuss the new subject of electromagnetism at all; thus he seems to have been ignorant of A.M. Ampère's contour theorem of 1826.

Green assumed that electricity and magnetism required one fluid each, and he interpreted their 'density' at a point as their quantity there (for example, p. 55). Apart from simple cases, he determined it from potential functions and equations such as (3).

In the second Part of the book Green considered various cases of electricity. The first one was the 'Leyden phial', both on its own and with several deployed 'in cascade'. Assuming from experimental evidence that exterior and interior potential were each constant, he linked them by Taylor's series truncated to the first term across the thickness of the glass, within which Laplace's equation obtained. At one point Green modified this equation by allowing for the curvature of the equipotential surface (p. 45), so that it became

$$\frac{d^2\bar{V}}{d\omega^2} = 4\pi\bar{\rho}\left(\frac{1}{R} + \frac{1}{R'}\right), \tag{4}$$

where $\bar{\rho}$ was the surface density at a point and R and R' the principal radii of curvature there. Perhaps he had read Gaspard Monge or some follower on differential geometry, though Lacroix's treatise might have provided enough information. Green's use of (4) has been little noticed; the equation was rediscovered on various later occasions [McAllister and Pedersen, 1988].

A variant upon Poisson was to consider two electrified spheres joined by 'an infinitely fine wire', when some simple forms for potential functions were found (pp. 55–56); a harder case was taken with a 'very thin spherical shell, in which there is a small circular orifice', which could be considered as planar when simplifying the form of the integral to show that small density would obtain there (pp. 57–61). Green then analysed phenomena involving 'long metallic wires, insulated and suspended in the atmosphere'; for example, when joining two spheres (pp. 63–65).

In all these cases Green presumed electrical conduction to be perfect; next he allowed for imperfection, which he likened to the effect of friction in mechanics (p. 70). His main example concerned the production of magnetism within a rotating body, where he drew upon [Barlow, 1825] for data (p. 75).

Green passed on to magnetism proper for the last Part of his book. He first determined potential for 'a very small body' (p. 83), perhaps like the monopole which had served a major role in Poisson's analysis; again Legendre functions were used. He then imitated Poisson in studying the 'magnetic state' of any body, and for once he used his theorem (1) to convert the potential function to a surface integral (p. 89). His last case dealt with magnetism in 'cylindric wires', where differential operators and complex variables led him to a closed form for density from which he could calculate values to set against the experimental data found by C.A. Coulomb as recorded in J.B. Biot's *Traité de physique* (1816) (pp. 111–115). The French influence remained central.

5 GREEN'S LATER RESEARCHES

All congratulations to Green for research and development; but his capacity for marketing and publicity was infinitesimal. The use of a subscription list was then becoming somewhat old-fashioned, as the market for science books and journals in Britain had grown significantly, albeit not to Parisian levels. He does not seem to have sent copies to leading scientists: for example, James Ivory (1765–1842), a leading student of equipotential surfaces who had disputed on precisely this topic recently with Poisson in the *Philosophical magazine.* Nor did he summarise its main results in that journal, or the *Quarterly journal*

of science, who surely would have taken a piece. Although the title page shows that copies were available from a few booksellers in London and Cambridge as well as in Nottingham, it is very unlikely that many customers parted with the requested 7 shillings and sixpence.

However, around 1830 Green did start to develop a career. After inheriting the mill from his father he appointed a manager, so giving himself more time for other activities. He made contact with one of his subscribers, the local scientific dignitary and mathematician Sir Edward ffrench Bromhead (1789–1855), member with Charles Babbage and John Herschel of the Analytical Society at Cambridge in the mid 1810s. Bromhead not only encouraged Green's work but also helped secure for him a Cambridge education at his own old college, Gonville and Caius. After four years' study there Green took the Mathematical Tripos examination in 1837, coming fourth Wrangler (with J.J. Sylvester ranked second); during much of 1839–1840 he was a fellow at the college. Colleagues for parts of this period included two other outsider mathematicians: Robert Murphy, who knew of his book and seems to have been the first to cite it; and Matthew O'Brien. Presumably his desire for this sort of career had caused him not to marry a local lace-dresser, Jane Smith, with whom he had seven children between 1824 and 1840.

During the 1830s Green produced nine research papers, most of them published in the *Transactions* of the Cambridge Philosophical Society, to which Bromhead had introduced him. They constituted a distinguished contribution to mathematical physics, typical for its time in exploiting analogies between types of phenomena and theories. Even though they appeared in orthodox journals, they did not gain the measure of attention that they deserved, though it is known that he sent some offprints to C.G.J. Jacobi.

In his first study Green handled the equilibrium of fluids in terms of Legendre functions. In the text he gave a more explicit formulation of Dirichlet's principle, and included the solution of the differential equation by a method now known as 'WKB', the initials of the surnames of three independent rediscoverers in 1926. He also considered the effect of the surrounding air on the motion of the pendulum, and in two other papers he handled the motion of waves in canals. In one of these he put forward the Dirichlet principle, which was to assume great status in potential theory when J.P.G. Lejeune-Dirichlet (1805–1859) began to use it from around the same time. When examining 'the reflection and refraction of sound', he gave the first detailed study of total internal reflection. In two papers on optics he followed the tradition of taking the phenomena as occurring in the elastic aether, and considered effects in both non-crystalline and crystalline media; some of his results related to work by Cauchy.

However, in all this writing Green only mentioned his book twice [1871, 120, 171], and gave no indication of its importance. After his death in 1841 it seems to have disappeared from sight and mind.

6 THE DISCOVERY OF THE BOOK

However, one little touch of publicity saved Green's book. He had given a few copies of it to the Cambridge coach Gowland Hopkins, who passed on a couple to his student William Thomson (1824–1907) in 1845. The young man recognised the importance of the book very quickly, and soon afterwards he shared it with colleagues in Paris. One of them was

Joseph Liouville, who also edited a mathematical journal; Thomson arranged for a reprint of the book not there but in the *Journal für die reine und angewandte Mathematik* run by A.L. Crelle, where the three Parts appeared in 1850, 1852 and 1854.

Despite this languid reappearance the word soon spread (the main results were of course in the first Part), and the theorem and function became standard concerns for potential theorists. When Green's book became well known it was clear that he had brought in a new phase the branch of mathematics that became known after him as 'potential theory'. Prior to the book various figures, especially Laplace and Lagrange and a few special results due to C.F. Gauss, had produced results usually concerning equipotential surfaces, extending those for potentials at points on principal axes of a body as established by Isaac Newton. Poisson had gone somewhat further, and more than he realised, with his 'simplifying' theorem. Now after Green's book the subject possessed new ways to handle many kinds of phenomena in mechanics and physics involving continuous bodies [Bacharach, 1883]. From the mathematical point of view surface integrals rose substantially in status from previous obscurity (and similarly for planar cases, line integrals were enhanced); one important consequence was the devising of further theorems of the kind of (1) [Cross, 1985], such as Stokes's (which is due to Thomson: §58.2).

Thomson also played an important role as researcher. In particular, he was inspired by the book to develop his 'method of images' for determining potentials; and when he and P.G. Tait produced their famous *Treatise on natural philosophy* (§40) they discussed (1) and included an extensive account of potential theory and Green's work, in the context of statics [Thomson and Tait, 1879, ch. 1, app. B; ch. 6, arts. 482–550]. However, progress was not straightforward: in particular, in 1870 Karl Weierstrass showed by means of a counter-example that the proof of Dirichlet principle was defective, for the optimal value that it located was not necessarily the one desired. This finding led to a huge complication of methods in potential theory.

An edition of Green's works was produced in 1871, and was reprinted in Paris thirty years later. The book itself was reprinted in Germany in 1890, and appeared in German translation five years later. This was a lot of publicity; of writings from the 1820s only Fourier's book on heat diffusion seems to have enjoyed a similar measure during that period. The transformation of classical mathematical physics into quantum physics and relativity theory in the 20th century did not eclipse his contributions; on the contrary, they gained new life. Green's theorem (a name due to Thomson and Tait) and function (Bernhard Riemann and Carl Neumann) have remained staple diet for applied mathematics until our time, and will doubtless last.

In the last 30 years aspects of Green's life and career have been rehabilitated. The biography [Cannell, 1993] formed part of a string of activities in Nottingham started in 1973 to restore the mill to working order; a science centre was installed there in 1985. A special honour was made in the bicentenary year, 1993; conferences were held in Nottingham and London, and after the latter event a plaque was unveiled in Green's memory in the floor of the nave of Westminster Abbey, close to the tomb of Isaac Newton and to the plaques for Michael Faraday, Clerk Maxwell, and his first publicist William Thomson, Lord Kelvin.

BIBLIOGRAPHY

Bacharach, M. 1883. *Abriss der Geschichte der Potentialtheorie*, Würzburg: Thein.

Barlow, P. 1825. 'On the temporary magnetic effect induced in iron bodies by rotation', *Philosophical transactions of the Royal Society of London*, 317–327.

Cannell, M. 1993. *George Green. Mathematician and physicist 1793–1841*, London: Athlone Press. [2nd ed. Philadelphia: SIAM, 2001.]

Cannell, M. 1999. 'George Green: An enigmatic mathematician', *American mathematical monthly, 106*, 136–151.

Cross, J.J. 1985. 'Integral theorems in Cambridge mathematical physics, 1830–55', in P.M. Harman (ed.), *Wranglers and physicists*, Manchester: Manchester University Press, 112–148.

Fourier, J.B.J. 1818. 'Note relative aux vibrations des surfaces élastiques . . .', *Bulletin des sciences, par la Société Philomatique de Paris*, 129–136. [Repr. in *Oeuvres*, vol. 2, 255–265.]

Grattan-Guinness, I. 1990. *Convolutions in French mathematics, 1800–1840*, 3 vols., Basel: Birkhäuser; Berlin: Deutscher Verlag der Wissenschaften.

Grattan-Guinness, I. 1995. 'Why did George Green write his essay of 1828 on electricity and magnetism?', *American mathematical monthly, 52*, 387–396.

Green, G. 1871. *Mathematical papers* (ed. N.M. Ferrers), London: Macmillan. [Repr. Paris: Hermann, 1903; New York: Chelsea, 1970.]

Green, H. 1946. 'A biography of George Green', in A. Montagu (ed.), *Studies and essays in* [. . .] *honor of George Sarton*, New York: Schuman, 545–594.

McAllister, I.W. and Pedersen, A. 1988. 'Green's differential equation and electrostatic fields', *Journal of physics, D 25*, 1823–1825.

Poisson, S.-D. 1824. 'Memoir on the theory of magnetism', *Quarterly journal of science, 17*, 317–334. [Translation of a summary of [Poisson, 1826], publ. in *Annales de chimie et de physique, (2) 25* (1824), 113–137, 221–223.]

Poisson, S.-D. 1826. 'Mémoire sur la théorie du magnétisme', *Mémoires de l'Académie des Sciences, 5* (1821–22), 247–338.

Poisson, S.-D. 1827. 'Mémoire sur la théorie du magnétisme en mouvement', *Ibidem, 6* (1823), 441–570.

Tazzioli, R. 2001. 'Green's function in some contributions of 19th-century mathematicians', *Historia mathematica, 28*, 232–252.

Thomson, W. and Tait, P.G. 1879. *Treatise on natural philosophy*, 2nd ed., 2 vols., Cambridge: Cambridge University Press.

Whitrow, G.J. 1984. 'George Green (1793–1841): a pioneer of modern mathematical physics and its methodology', *Annali dell'Istituto di Storia della Scienza di Firenze, 9*, no. 2, 47–68.

Whittaker, E.T. 1951. *History of the theories of aether and electricity. The classical theories*, London: Nelson.

CHAPTER 31

C.G.J. JACOBI, BOOK ON ELLIPTIC FUNCTIONS (1829)

Roger Cooke

This treatise was the first systematic exposition of elliptic functions using the new techniques made available by the theory of analytic functions of a complex variable. Niels Henrik Abel was also an important pioneer.

First publication. *Fundamenta nova theoriae functionum ellipticarum*, Königsberg: Borntraeger, 1829. 192 pages.

Reprint. In *Werke*, vol. 1, Berlin: Reimer, 1881 [repr. New York: Chelsea, 1969], 49–239.

Related articles: Euler *Introductio* (§13), Cauchy on complex-variable analysis (§28).

1 ELLIPTIC INTEGRALS

The success of the calculus in solving physical problems through the formulation and solution of differential equations gave a high value to the computation of integrals. Very early on, a wide variety of integrals that can be expressed by elementary functions were discovered. The treatment of such integrals was unsystematic, however, due to the absence of the concept of an inverse function. For example, the relation that would nowadays be expressed as the indefinite integral

$$\int \frac{dx}{\sqrt{2x - x^2}} = \arccos(1 - x) + C \tag{1}$$

was written by G.W. Leibniz as the equation

$$a = \int dx : \sqrt{2x - xx}, \tag{2}$$

where x is the versed sine of a. That is, $x = 1 - \cos(a)$. This way of expressing the integral points up an important feature of the modes of thought of mathematicians up to the end

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

of the 18th century. They recognized variables and relations between them, but the strict separation between independent and dependent variables was not clear. The functions we now call the inverse trigonometric functions were not an object of study in themselves. Another obvious feature missing from (2), namely the limits of integration in the definite integral, was introduced (by Joseph Fourier) only in the 1810s.

The mathematical treatment of integrals involving the square root of a quadratic polynomial was greatly simplified by the familiarity of the trigonometric functions. The success in this area often served as a guide in the study of the more complicated integrals known as elliptic integrals, and even today provides a framework for understanding the issues involved in the historical development of the topic. For that reason, we shall begin with a brief discussion of the elementary integrals leading to transcendental functions.

To obtain a complete understanding of integrals whose worst irrationality is the square root of a quadratic polynomial only two problems needed to be solved. The first was to classify all the essentially distinct kinds of integrals of the form

$$\int R\bigl(x, \sqrt{ax^2 + bx + c}\bigr)\, dx, \tag{3}$$

where $R(x, y)$ is a rational function of x and y. The second was to eliminate restrictions on the domain of the integrand by allowing complex values. The first step was easily taken. Merely completing the square and then making simple linear substitutions made it possible to restrict consideration to only three different kinds of quadratics, namely $1 - x^2$, $x^2 - 1$, and $1 + x^2$. The second step, which was taken only piecemeal until the 19th century, made it possible to reduce all three of these cases to one case.

One important place where integrals of this type arise is in the rectification of a circle. Given the circle whose equation is $x^2 + y^2 = 1$, the length of arc from the point $(0, 1)$ to the point (x, y) is given by integrating the differential of the arc length, denoted ds and given by

$$ds = \sqrt{1 + \left(\frac{dy}{dx}\right)^2}\, dx = \frac{dx}{\sqrt{1 - x^2}}. \tag{4}$$

The classical problem of bisecting or trisecting an angle is equivalent to the same problem for arcs of a circle. For that reason, a problem of fundamental geometric significance was the problem of dividing an arc into n equal pieces. The solution of this problem depends on the formula $\cos\theta + i \sin\theta = e^{i\theta}$. It shows, in particular, that the real part of the nth root of $e^{i\theta}$ is $\cos(\theta/n)$, so that expressing the trigonometric functions of θ/n in terms of those same functions of θ becomes a problem having an interesting mathematical application. In particular, applying the binomial theorem to the equation $e^{i\theta} = (e^{i\theta/n})^n$ shows that $\cos(\theta/n)$ is an algebraic function of $\cos\theta$. For example, $\cos\theta = 4(\cos(\theta/3))^3 - 3\cos(\theta/3)$. The particular form of the algebraic function makes it possible to construct certain regular polygons with straightedge and compass (for example, the heptakaidecagon, or regular 17-sided rectilinear figure) and to prove that others (for example, the nine-sided nonagon) cannot be so constructed. The possibility of constructing the heptakaidecagon was discovered by C.F. Gauss in 1796, and the impossibility of constructing the nonagon was proved by Pierre Laurent Wantzel in 1837.

The periodicity of the trigonometric functions plays an important role in the division problem. Once it is known that there is a polynomial $p(x, y) = q(x) - y$ of degree 3 in x (for example) such that $p(\cos(\theta/3), \cos\theta) \equiv 0$, the fact that $\cos(\theta + 2\pi) = \cos\theta$ shows that

$$p\left(\cos\left(\frac{\theta + 2m\pi}{3}\right), \cos\theta\right) \equiv 0 \tag{5}$$

for all integers m. It follows that the roots of the equation $p(x, \cos\theta) = 0$ must be $\cos((\theta + 2m\pi)/3)$ for $m = 0, 1$, and 2. Therefore it is a simple matter to construct these roots. Although the division problem for elliptic integrals is more complicated, the ultimate solution of the division problem for these integrals was to be based on the double periodicity of the inverse functions, exactly as in (5).

Because of the importance of these relations between functions one of whose arguments is a multiple of the argument of the other, the division problem became an important topic in the study of integrals of algebraic functions. Since the variable s in (4) is simply $\arcsin(x)$, obviously the variable $z = ns = n\arcsin(x)$ will satisfy $dz = n\,ds$, so that, if $z = \arcsin(y)$, then

$$\frac{dy}{\sqrt{1 - y^2}} = dz = \frac{n\,dx}{\sqrt{1 - x^2}}. \tag{6}$$

This last relation is the differential equation satisfied by the function $y = \sin(n\theta)$ in terms of the independent variable $x = \sin\theta$. Differential equations of this type were the avenue of approach to the multiplication/division problem for general algebraic integrals. Leonhard Euler, in particular, was impressed by the fact that the relation between x and y expressed by this equation was an *algebraic* relation, even though neither indefinite integral was an algebraic function of its argument. One can see that this general division/multiplication problem can be solved using the law of addition for arguments in the trigonometric functions, and these were well known by the time calculus was invented. The problem of dividing an arc into any number of equal pieces was thus seen to be a matter of solving a polynomial equation, although the equation could not always be solved using only straightedge and compass, or even using only a finite number of arithmetic operations and root extractions.

This brief discussion of elementary integrals highlights three important questions that would naturally be asked when more complicated integrals came to be considered: a) How many essentially distinct integrals of a given type exist? b) How are such integrals to be multiplied and divided by integers? (In other words, how is the analog of (6) to be solved?) c) What are the analogs of the trigonometric functions for such integrals? The answer to the third question would of course provide the answer to the second.

A large number of examples of integrals involving the square root of a quadratic were incorporated in Euler's treatise *Introductio in analysin infinitorum* (1748) (§13) and in his *Institutiones calculi integralis* (three volumes, 1768–1770). Such integrals arise in very simple physical problems, such as the motion of a harmonic oscillator (an undamped, freely vibrating spring, for example), described by the differential equation $y'' + \omega^2 y = 0$. Since the independent variable does not appear explicitly, the substitution $p = y'$, $y'' = p\frac{dp}{dy}$,

leads to

$$p\,dp + \omega^2 y\,dy = 0, \text{ so that } p^2 + \omega^2 y^2 = C^2, \text{ which yields } x = \int \frac{dy}{\sqrt{C^2 - \omega^2 y^2}}. \quad (7)$$

Here both mathematical and physical considerations suggest that the integral is somehow 'backwards', that one should be regarding x, rather than y, as the independent variable. That is to say, they confirm what Leibniz already knew: that this equation asserts that y is a trigonometric function of x. The mathematical difficulty is that the integrand has a singularity at the points $y = \pm C/\omega$. In order for x and y to have a physical interpretation, y must be confined to the finite interval between these two values. Mathematically, however, nothing prevents y from being outside that range. The physical difficulty is that x represents time, which should always be an independent variable. The physical difficulty can be handled by writing

$$y = \frac{C}{\omega}\sin(\omega x - K) \text{ instead of } x = \frac{1}{\omega}\big(\arcsin(\omega y/C) + K\big). \quad (8)$$

The mathematical question, however, remains. In the 18th century Euler began using imaginary substitutions to handle problems of this sort. In the present case the substitution $y = iu$, makes it possible to express x as $\frac{i}{\omega}\log(\omega u/C + \sqrt{1 + (\omega^2 u^2/C^2)})$. The need for imaginary quantities, which naturally arises when square roots occur in an integrand, shows that a complete understanding of such integrals requires the theory of analytic functions of a complex variable. Until this theory was developed in the early 19th century by C.F. Gauss and especially A.L. Cauchy (§28), the full understanding of even these elementary integrals was delayed.

Euler was an analyst par excellence, who manipulated symbolic expressions with great ease. But the creation of the proper language for discussing algebraic integrals of this type required something more than the introduction of symbols for imaginary quantities. That extra ingredient was the geometric interpretation of complex numbers as the Euclidean plane and the definition of integrals over paths in the complex plane. That realization came to Gauss and others in the early 19th century. In the meantime, a number of physical and geometric problems led to still more complicated algebraic integrals. For example, the equation of motion of a frictionless oscillating pendulum, namely

$$y'' + \omega^2 \sin y = 0 \quad (9)$$

can be handled just like the equation of the vibrating spring discussed above, leading to the equation

$$dx = \frac{dz}{\sqrt{C^2 - \omega^2 \sin^2 z}} = \frac{1}{C}\frac{dz}{\sqrt{1 - k^2 \sin^2 z}}, \quad (10)$$

where $z = y/2$ and $k = \omega/C$. The substitution $u = \sin(z)$ changes this differential into

$$dx = \frac{1}{C}\frac{du}{\sqrt{(1 - k^2u^2)(1 - u^2)}}. \quad (11)$$

The integrands that appear in (10) and (11) are the standard forms of what is now called the *elliptic integral of first kind*, a name bestowed by A.M. Legendre (1752–1833), who classified all integrals involving at worst the square root of a cubic or quartic polynomial into three distinct types of elliptic integrals. The word *elliptic* arose because the rectification of the ellipse leads to such an integral (of second kind). Elliptic integrals arise in the study of the rotation of a rigid body, and a complete understanding of the special cases of this motion studied by Euler and Lagrange requires an analysis of such integrals.

2 ELLIPTIC INTEGRALS FROM FAGNANO TO LEGENDRE

2.1 Fagnano and Euler

The first person to have considered the problem analogous to division and multiplication for more general integrals seems to have been Count Giulio Carlo de' Toschi di Fagnano (1682–1766), a diplomat and amateur mathematician who worked on the problem of rectifying complicated curves. He found a number of cases in which the difference of two arcs of a conic section could be expressed as an elementary integral resembling (11). He then extended this work to higher-order curves such as the lemniscate, whose equation is $(x^2+y^2)^2 = 2(x^2-y^2)$, or, in polar coordinates, $r^2 = 2\cos(2\theta)$. The element of arc of such a curve is given in polar coordinates by

$$ds = \sqrt{2\sec(2\theta)}\,d\theta = \sqrt{2}\,d\theta/(1-2\sin^2\theta)^{1/2}. \tag{12}$$

The substitution $u = \tan\theta$ makes the element of arc $ds = \sqrt{2}\,du/\sqrt{1-u^4}$. The division problem naturally arises for this curve, just as it does for a circle, and that means solving the analog of (6). In a 1718 paper entitled 'Metodo per misurare la lemniscata' Fagnano was able to solve the differential equation

$$\frac{dz}{\sqrt{1-z^4}} = \frac{2\,du}{\sqrt{1-u^4}} \quad \text{in the form} \quad \frac{u\sqrt{2}}{\sqrt{1-u^4}} = \frac{1}{z}\sqrt{1-\sqrt{1-z^4}} \tag{13}$$

and thereby find an algebraic solution of the problem of doubling or bisecting an arc of the lemniscate. By such methods he was able to show how to divide a quadrant of the lemniscate into various numbers of equal parts, of the form $2^m p$, where $p = 3$ or $p = 5$, and also of the form 2×3^m.

In 1750 the now elderly Fagnano sent his collected mathematical works to the Berlin Academy, hoping for election to membership. The person called upon to judge them was Euler, who realized the importance of what Fagnano had done and in [Euler, 1761] noted the analogy with the corresponding equation that results when the fourth powers are replaced by squares, and in particular the fact that the relation between x and y was once again algebraic, even though in this case the indefinite integrals were not only not algebraic, but not even elementary transcendental functions. Apparently having proceeded by inspired guessing, Euler gave the general solution of this differential equation in the case $m = n = 1$ as

$$x^2 + y^2 + c^2x^2y^2 = c^2 + 2xy\sqrt{1-c^4}. \tag{14}$$

Thus, even though the individual integrals are not algebraic, or even elementary transcendental functions, the general solution of the differential equation is of the form $P(x, y; c) = 0$, where P is a polynomial in its three arguments. If $\psi(x)$ denotes an indefinite integral of the function $1/\sqrt{1-x^4}$, then this function is neither algebraic nor an elementary transcendental function. Nevertheless the relations $\psi(x) + \psi(y) = K$ and $P(x, y; c) = 0$ are equivalent. Euler showed in fact that the expression $1 - x^4$ could be replaced by a completely arbitrary fourth-degree polynomial. The general integral of the equation would still be a polynomial in x and y containing an arbitrary parameter. Euler remarked that this property would cease to hold when polynomials of higher degree were allowed, since the general integral of the reciprocal of the cubic polynomial $1 + x^3$ (which is the square root of a sixth-degree polynomial) contains terms involving the arctangent and also terms involving the logarithm, and there is no algebraic relationship between these two functions. Here we find Euler at the very edge of applicability of his techniques. He knew a considerable amount about the logarithm of a complex number, in particular that its real part is a logarithm and its imaginary part an arctangent, but he apparently did not realize that the introduction of complex variables would have allowed him to express this integral as a simple logarithm.

Through this general solution Euler made a connection between differential equations of this type and algebraic addition theorems for the integrals. Specifically, he noted that the equation could be written as

$$\int \frac{dx}{\sqrt{X}} + \int \frac{dy}{\sqrt{Y}} = \int \frac{db}{\sqrt{B}}, \tag{15}$$

where b is a constant. This remark was to play an important role in the future development of the subject, leading Legendre to the addition formula for elliptic functions and Abel to the discovery of a general theorem guaranteeing the validity of an equation of this type in which the limits of integration on the right are algebraic functions of those on the left.

2.2 *Legendre*

The form of elliptic integrals that is nowadays best known was adapted by Jacobi from Legendre. Pursuing the first of the tasks noted above, Legendre sought to determine the number of essentially distinct integrals one could generate from formulas of the form

$$\int R\big(x, \sqrt{ax^4 + bx^3 + cx^2 + dx + e}\big)\, dx. \tag{16}$$

Legendre's result was that such integrals were of three basic types, which he called first, second, and third kinds. Each integral contained an unspecified constant, which he called its *modulus*. The integral of third kind contained a second parameter in addition to the modulus, which he called the *parameter*. By a trigonometric substitution he was able to present the general elliptic integral of first kind in the form

$$F(\phi) = \int \frac{1}{\Delta} d\phi, \quad \text{where } \Delta = \sqrt{1 - c^2 \sin^2 \phi}. \tag{17}$$

Using the function F of (17), he noted that the solution of the differential equation

$$\frac{d\phi}{\sqrt{1-c^2\sin^2\phi}}+\frac{d\psi}{\sqrt{1-c^2\sin^2\psi}}=0 \tag{18}$$

could be written in the form $F(\phi)+F(\psi)=F(\mu)$, where μ is a constant. He gave its general solution, obtained, as he said, 'by Euler's method', in the form

$$\cos\phi\cos\psi-\sin\phi\sin\psi\sqrt{1-c^2\sin^2\mu}=\cos\mu. \tag{19}$$

From this last relation Legendre derived the addition formula for elliptic functions of first kind. To be specific, he showed that if $F(\phi)+F(\psi)=F(\mu)$, then

$$\sin\mu=\frac{\sin(\phi)\cos(\psi)\Delta(\psi)+\cos(\phi)\sin(\psi)\Delta(\phi)}{1-c^2\sin^2\phi\sin^2\psi}. \tag{20}$$

As was mentioned above in connection with elementary integrals, having an algebraic addition formula of this type provides a complete solution of the multiplication/division problem for such integrals, and Legendre was quick to point out this fact. He used the notation $\psi=\phi_n$ to denote the solution of the differential equation

$$\frac{n\,d\phi}{\sqrt{1-c^2\sin^2\phi}}=\frac{d\psi}{\sqrt{1-c^2\sin^2\psi}} \tag{21}$$

such that $\phi_n=0$ when $\phi=0$. He showed, for example, that

$$\sin\phi_3=\frac{3\sin\phi-4(1+c^2)\sin^3\phi+6c^2\sin^5\phi-c^4\sin^9\phi}{1-6c^2\sin^4\phi+4c^2(1+c^2)\sin^6\phi-3c^4\sin^8\phi}. \tag{22}$$

The addition formula gives an algebraic solution of the problem of dividing an elliptic integral of first kind into equal pieces. In general, as Legendre showed, the equation for $\sin\phi$ in terms of $\sin\phi_n$ is of degree n^2. The analogy with the trigonometric integrals was now apparent, and Legendre accordingly gave the variable ϕ the name of *amplitude*. In writing the equation (20), in which the function F is stripped away, he was effectively regarding the elliptic integral as the function inverse to the amplitude. However, while he certainly recognized the amplitude as a variable, he did not think of it as being a 'dependent' variable determined by the value of the integral. The explicit recognition of that relationship, first made by N.H. Abel, was of great importance in the subsequent development of the subject. Legendre's work also could have benefited from a more generous use of complex variables. He allowed the parameter in an elliptic integral of third kind to be imaginary, but did not explore the possibility that the amplitude ϕ might be complex. Since ϕ is interpreted as the upper limit of an integral, doing so would have meant integrating over complex paths. That could have been done in the mid 1820s, as Cauchy had defined such integrals (§28). Although Jacobi realized what a great liberation this step meant, neither Abel nor Jacobi used contour integrals as we now know them. Their use of complex variables during the 1820s

tended to be formal, glossing over certain subtleties, such as the different branches of the square root, for example. Integrating over contours would have forced them to clarify such matters.

2.3 Transformation of elliptic functions

In classifying elliptic integrals into the minimum number of distinct kinds, Legendre made a remarkable discovery connecting two elliptic integrals with different moduli. For integrals of first kind this discovery can be written as

$$F(c', \phi') = \frac{1+c}{2} F(c, \phi), \tag{23}$$

where

$$\begin{aligned} &c' = 2\sqrt{c}/(1+c), \text{ that is, } c = \frac{2-(c')^2 - 2\sqrt{1-(c')^2}}{(c')^2}, \quad \text{and} \\ &\sin(2\phi' - \phi) = c \sin\phi. \end{aligned} \tag{24}$$

By iterating the transformations in these last equations, he produced an entire scale of algebraic relations among elliptic integrals with different arguments. It thus appeared that these integrals were much richer in symmetries than the trigonometric functions and their inverses. Legendre did not yet know how much richer. He valued the relation (23) especially because of its computational implications. Starting with c' and iterating the procedure a finite number of times made it possible to express the function $F(c', \phi')$ in terms of an integral where the modulus is as small as desired. But when the modulus is negligible, the integral is approximately equal to its upper limit. Thus, transformation of the integral makes it possible to approximate it by a multiple of the upper limit of the transformed integral, and both the modulus and the upper limit of the transformed integral are computable as elementary functions of the original modulus and upper limit. Legendre called the doubly infinite sequence of moduli and functions generated by iterating the operation (23) in both directions a *scale* (*échelle*).

3 GENESIS OF THE *FUNDAMENTA*

3.1 Legendre's treatise

Legendre published his results on elliptic integrals in his *Exercices de calcul intégral* (1811). In the early 1820s he decided to write an extensive treatise on elliptic functions, giving the main mathematical results and applications in the first volume and tables of the values of these functions in the second volume. In 1825, just as he was about to complete his treatise, he noticed yet a second scale of transformations, namely

$$F(\alpha, \omega) = m F(c, \phi), \tag{25}$$

where

$$1-\alpha^2\sin^2\omega=(1-c^2\sin^2\phi)\left(\frac{1-(k/m)\sin^2\phi}{1+k\sin^2\phi}\right)^2, \tag{26}$$

$k=\frac{1}{4}(m-1)(m+3)$, $1<m<3$, and the moduli c and α are given by

$$c^2=\frac{(m-1)^3(1+3/m)}{16} \quad \text{and} \quad \alpha^2=\frac{(m-1)(1+3/m)^2}{16}. \tag{27}$$

Legendre was very pleased with this result, even though it meant that he would have to rewrite much of his nearly completed treatise. The scale of moduli that resulted from this transformation converged to zero much faster than those in the scale he had discovered previously, making a more rapid approximation of the integral possible. Moreover, the parameter m that defined each transformation could be obtained from the initial modulus c by solving a fourth-degree equation. He also noticed that the substitution $m\mapsto 3/m$ resulted in $c^2\mapsto 1-\alpha^2$ and $\alpha^2\mapsto 1-c^2$. In other words, each modulus was replaced by what he called the 'complement' of the other. Lastly, he noticed that two different scales, starting from different moduli and amplitudes, would stand in the relation

$$\frac{F(c_r,\phi_r)}{F(b_r,\psi_r)}=3\frac{F(c,\phi)}{F(b,\psi)}. \tag{28}$$

What is remarkable, considering that he noticed relation (28), is that Legendre apparently did *not* notice the connection this second scale of transformations has with the division problem for integrals of first kind. He had derived the scale by beginning with the transformation

$$\sin\omega=\sin\phi\frac{m+h\sin^2\phi}{1+k\sin^2\phi} \tag{29}$$

and later taking $h=((m-1)/2)^2$. He thought that the restriction $1<m<3$ was necessary in order to keep the moduli between 0 and 1. Evidently, he did not notice that the range $-3<m<-1$ would have the same effect, that the amplitude ω would still be an increasing function of ϕ when $m<0$ if $\sin\phi$ were replaced by $-\sin\phi$ in relation (26), and that the result would be the relation

$$\int_0^{\omega}\frac{dt}{\sqrt{1-\alpha^2\sin^2 t}}=|m|\int_0^{\phi}\frac{dt}{\sqrt{1-c^2\sin^2 t}}. \tag{30}$$

The substitution $m\mapsto -3/m$ in a second application of the transformation between ω and a new variable θ would then exactly reverse the roles of the two moduli, resulting in an algebraic relation between ϕ and θ equivalent to the equation

$$\int_0^{\theta}\frac{dt}{\sqrt{1-c^2\sin^2 t}}=3\int_0^{\phi}\frac{dt}{\sqrt{1-c^2\sin^2 t}}. \tag{31}$$

Legendre no doubt believed that he had solved the division problem by producing the addition theorem (20) for the amplitudes; thus, he probably was not looking at this aspect of his new scale.

3.2 Jacobi's early discoveries

Legendre's treatise bears a nominal date of 1825, but the first volume did not actually appear until January 1827. Eight months after it appeared Legendre was astounded to find that Jacobi had discovered this second scale of transformations, and apparently many others besides.

In a letter written to editor H. Schumacher on 13 June and published in the *Astronomische Nachrichten* in September 1827, Jacobi gave the formula (24), using homogeneous coordinates in place of m and k so as to reduce the number of radicals, and he pointed out the connection between the transformation problem and the division problem expressed by (26). As he noted, this transformation made it possible to solve the equation (26), which amounts to the ninth-degree equation (22), by purely algebraic operations. First one finds m in terms of c (a fourth-degree equation), then solves the equation (24). Replacing m by $-3/m$ and solving (24) again (with new variables) then leads to (26), thereby making the trisection problem theoretically solvable by a finite number of algebraic operations. There is no doubt that Jacobi had both discovered Legendre's second scale independently of Legendre and gone a step further, producing a similar transformation that divides the elliptic integral by 5.

In a letter to Legendre dated 12 April 1828 Jacobi described the route by which he had obtained these results. He used the differential form (11) rather than (10), and asked when it was possible to obtain an equation of the form

$$\frac{dy}{\sqrt{(1-\alpha y)(1-\alpha' y)(1-\alpha'' y)(1-\alpha''' y)}} = \frac{dx}{M\sqrt{(1-\beta x)(1-\beta' x)(1-\beta'' x)(1-\beta''' x)}} \tag{32}$$

by a transformation of the form $y = U/V$, where U and V are relatively prime polynomials in x of degree at most n, where n is an odd integer. Here the parameters containing α are regarded as given, while M and the βs are regarded as adjustable so as to obtain the equation (32). In March 1827 he had worked out that U and V must satisfy an equation of the form

$$\begin{aligned}&(V-\alpha U)(V-\alpha' U)(V-\alpha'' U)(V-\alpha''' U)\\ &\quad = M(1-\beta x)(1-\beta' x)(1-\beta'' x)(1-\beta''' x)(1-\gamma_1 x)^2\cdots(1-\gamma_{2n-2}x)^2,\end{aligned} \tag{33}$$

which contains a total of $4n+2$ adjustable parameters (in addition to the $2n-1$ coefficients of U and V that can be chosen independently, there are $2n+3$ adjustable parameters on the right-hand side of (33)). The squared factors on the right actually constitute the polynomial $\{U\frac{dV}{dx} - V\frac{dU}{dx}\}^2$, as Jacobi showed easily. The equations that have to be satisfied break up into four sets, and from them Jacobi was able to work out the transformation for $n = 3$. Looking long and hard at the form of this transformation, he conjectured a general

form for the second modulus that would hold for any integer. When he tested his conjecture with $n = 5$, he found that it worked. These two cases formed the content of his first letter to Schumacher, sent on 13 June. Apparently, he wanted to work out a general proof that his formula would work for any integer n, and so he waited until August to send a statement of his conjecture to Schumacher and to Legendre. These letters, as he admitted, were disingenuous, since he couched his conjecture as a theorem. He was sure it was true, but had no proof as yet. As he stated in the 12 April letter [*Works*, vol. 1, 415–416],

> You expressed the wish that I would give the train of ideas that led me to my theorems. However, the route that I followed was not mathematically rigorous. Once the discovery was made, it was possible to replace it by a different route leading rigorously to the same results. Thus, I write what follows in confidence. The first discovery that I made (in March 1827) was the equation $T/M = V\,dU/dx - U\,dV/dx$. From that equation I realized that the problem was a *determinate* problem of algebraic analysis for an arbitrary integer n, since the number of arbitrary constants equals the number of conditions. Using undetermined coefficients, I constructed the transformations corresponding to the numbers 3 and 5. Since the quartic equation that the first of these led me to had nearly the same form as the equation for trisection, I suspected some relation between the two. By a stroke of luck, I noticed the complementary transformation that leads to the multiplication in both of these cases. At that point, I wrote my first letter to M. Schumacher, since the method was general and had been verified by examples. Later, examining the two substitutions $z = (ay + by)/(1 + cy^2)$ and $y = (a'x + b'x^3)/(1 + c'x^2)$ in the form given in my first letter, I saw that when $x = \sin\mathrm{am}(2K/3)$, z must vanish, and since b/a is positive in the given form, I concluded that y must also vanish. In this way, I found the factorization by induction, which being confirmed by examples, I gave the general theorem in my second letter to Mr. Schumacher [...] Since all this was confirmed by examples, I made bold to send a first letter to you, which you received with such graciousness. The proofs were discovered only later.

Since Jacobi did not provide the general proof and Schumacher was not a specialist in this area, the latter appealed to Gauss for advice. Gauss, of course, recognized that the result was correct, since he had been in possession of similar results for two decades or more. However, he had been considering the writing of a treatise of his own on this topic and was annoyed that Schumacher had compromised his claim to priority by showing him Jacobi's results. Schumacher then pressed Jacobi to provide the proof [Ore, 1957, 185–188].

3.3 Abel's first memoir

It is likely that while working on the proof Jacobi chanced to see a memoir of Niels Henrik Abel, which appeared in Crelle's *Journal für die reine und angewandte Mathematik* as [Abel, 1827]. Abel derived Legendre's addition formula for elliptic integrals of first kind

in the form

$$\phi(\alpha+\beta)=\frac{\phi(\alpha)f(\beta)F(\beta)+\phi(\beta)f(\alpha)F(\alpha)}{1+e^2c^2\phi^2(\alpha)\phi^2(\beta)}, \tag{34}$$

where

$$\alpha=\int\frac{1}{\sqrt{(1-c^2x^2)(1+e^2x^2)}}dx, \tag{35}$$

$$f(\alpha)=\sqrt{1-c^2\phi^2(\alpha)},\quad F(\alpha)=\sqrt{1+e^2\phi^2(\alpha)},\quad\text{and}\quad x=\phi(\alpha). \tag{36}$$

(The integral in (35) goes from 0 to x.) It was this last device, regarding the limit of integration as a function of the value of the integral, that ultimately revealed why Jacobi's transformations work. Abel also was interested in the division problem, and he particularly wanted to do for the lemniscate what Gauss had done for the circle: prove that a straightedge-and-compass division of it into m equal arcs is possible for any power of 2 times a prime of the form 2^n+1. (It is easy to see that n must be zero or a power of 2 if this number is prime. Such numbers are called *Fermat primes*, and only 5 are known to exist.) He succeeded in doing so by allowing the variables to be complex numbers. In that way, he discovered the double periodicity of the inverse functions. As remarked above, this double periodicity made it possible to construct explicitly the roots of the polynomial that expressed $\phi(n\alpha)$ in terms of $\phi(\alpha)$. To take the case $n=3$ as typical, it was known (see (22)) that there are polynomials $p(x)$ and $q(x)$ of degrees 9 and 8 respectively such that the identity $p(\phi(\alpha))-\phi(3\alpha)q(\phi(\alpha))\equiv 0$ holds. If ω and ϖ are two independent basic periods of $\phi(\alpha)$, it follows that the nine roots of the equation $p(x)-\phi(3\alpha)q(x)=0$ must be $\phi(\alpha+(m\omega+m'\varpi)/3)$, where m and m' range from 0 to 2 independently.

If Jacobi looked at this paper, he could not have helped noticing this device. Of course, he might have discovered it independently. In any case, he certainly used the device in his proof, which he posted on 18 November and Schumacher published in December. This article contained the first introduction of the famous Jacobi elliptic functions. He wrote $x=\operatorname{sin am}\Xi$, when $\Xi=\int_0^x\frac{dx}{\sqrt{(1-x^2)(1-k^2x^2)}}$. To construct his transformation in terms of the roots of the polynomial involved, Jacobi made use of the fact that the function x has period $4K$ in order to prove that his conjectured formula for the transformed amplitude is correct.

Actually, from Jacobi's formula for Ξ, there is no easy way to allow this variable to exceed K. However, if $\phi=\operatorname{am}\Xi$ is defined first, using (17), it is easy to see that both ϕ and Ξ tend to infinity together, and each is a well-defined function of the other. Jacobi used this slight ambiguity in the meaning of Ξ without comment.

3.4 The rivals

In his third letter to Schumacher, published in December 1827, Jacobi said that he would give the proof 'assuming certain auxiliary considerations that have already been published in part elsewhere'. This phrase is conveniently ambiguous; it probably alludes to Abel's paper, but avoids mentioning it outright. Abel noticed Jacobi's work, and in a follow-up to

his two long papers in Crelle's *Journal*, he pointed out that Jacobi's formula followed from one of his. Jacobi was aware that his rival was formidable and that he could not afford to spend two years doing nothing but writing his treatise. While working on it, he published a steady stream of papers in Crelle's *Journal* containing most of the contents of the treatise, thereby establishing an uncontested claim to the discovery of theta functions, one of the handiest tools in algebraic function theory. The brilliance of Jacobi's mind showed itself at its finest when he was able to use the fact that a certain product of theta functions has no vanishing Taylor coefficients to prove Euler's famous theorem that every positive integer is the sum of at most four squares. The rivalry between the two young geniuses never developed; sadly, Abel died in 1829, aged 26.

As for Legendre, he was content to let the young ones work out their ideas, knowing he could not possibly keep up with them. In general, Jacobi's results were written more or less in Legendre's style and hence were easier for the old man to understand. He did not understand Abel's approach, and considered that his own role was to serve as the herald announcing their discoveries to the mathematical world. He performed that role well. Jacobi, however, understood Abel clearly, and was happy to translate what Abel wrote into language that Legendre could understand. He may have done so too well. In a letter to Legendre dated 12 January 1828 he made all too plain the connection between Abel's work and the new notation that he introduced in his own proof, saying that he needed to use certain formulas given first by Abel. Jacobi did not actually admit to getting this idea from Abel, but he also did not explicitly claim independent discovery of it. Legendre expressed his displeasure diplomatically, saying only that he regretted that Jacobi did not have the sole claim to the credit for the proof [Jacobi, *Works*, vol. 1, 420]:

> To establish the principle of your proof, you say, you must have recourse to the analytic formulas for multiplication *given for the first time by M. Abel*. This admission demonstrating your candor, a characteristic of true talent, pains me slightly; for, while giving due credit to M. Abel's beautiful work, ranking it, however, well below your own discoveries, I could wish that the credit for the latter, that is, their proofs, belonged entirely to you.

4 THE AUTHOR

It is time formally to meet our author. Carl Gustav Jacob Jacobi was born in Potsdam on 10 December 1804, the son of a prosperous Jewish banker named Simon Jacobi. His elder brother Moritz became an architect, physicist and engineer of note. Carl showed outstanding aptitude in languages and mathematics at an early age and completed the normal school course at the age of 12. Since he could not enter a university before the age of 16, he had four years to devote to absorbing yet more languages and his primary love, mathematics. Like his great contemporary and rival Abel, he made a profound study of Euler's *Introductio in analysin infinitorum* and worked on the problem of the quintic equation (§29). He entered the University of Berlin in 1821, at a time when it had not yet distinguished itself. Although he was only 17, his own independent study was more than sufficient preparation for the career he was to follow. By 1825, despite the prejudice against Jews—maybe exacerbated in his case by his own abrasive personality—he was teaching at the University

of Berlin. Klein, in his history of 19th-century mathematics, notes that Jacobi was the first Jewish professor in Germany [1926, 109]. It is true that Jacobi had converted to Christianity in 1825, but such conversions had not always been sufficient to overcome the prevailing antisemitism.

In 1827, as already noted, Jacobi sent Legendre a letter containing his discoveries in elliptic functions. Legendre's enthusiastic response and Jacobi's continuing first-rate research started a brilliant career. He was an inspiring lecturer, who received regular promotions and was made a full ('Ordinarius') professor in Königsberg by 1832. He enriched many areas of mathematics, including differential geometry, mechanics, differential equations, and number theory.

But three personal misfortunes clouded Jacobi's life somewhat. He lost his inherited wealth in a general depression; and in 1842 he developed diabetes, incurable and untreatable in his day. Finally, in the revolutionary year of 1848 he made an unwise political speech that generally alienated all parties. The result was a denial of his petition to teach at the University of Berlin and the revocation of the supplemental allowance that had made it possible for him to live there. The Prussian government eventually realized Jacobi's value for its own prestige and relented to the extent of allowing him to give lectures at the University of Berlin. However, weakened no doubt by diabetes, he contracted influenza in early 1851 and then smallpox, from which he died on 18 February.

5 CONTENTS OF THE WORK

The 185 pages of this treatise are summarised in Table 1. It was divided into 66 paragraphs, representing 27 topics grouped into two major chapters. The first chapter (Sections 1–34) discusses the transformation of elliptic functions, the second (Sections 35–66) is devoted to various series and product representations of them.

Jacobi explained his reasons for publishing this work as a treatise in his preface:

> About two years ago, wishing to study the theory of elliptic functions in more detail, I entered upon investigations of extreme importance, in that they seemed to give the theory an entirely new form and to provide a remarkable advance in the universal art of analysis. Having succeeded beyond my expectations, due to the difficulty of the matter, in bringing this research to a conclusion, I communicated it to the mathematical community with its main points in abbreviated form and without proof, then soon afterwards with the proofs added, since the latter were urgently to be desired and the results, being newly discovered, could not be taken on faith. At the same time I was eager to publish a systematic exposition of the investigations I had undertaken. It is to satisfy this desire at least in part that I have undertaken to publish the foundations upon which my investigations are based. We now commend these new foundations of the theory of elliptic functions to the indulgence of the mathematical community.

The first seven Sections of the work contain the exposition of the complete theory of transformation of elliptic integrals of first kind, together with examples and tables of the various transformations. The examples of division and multiplication that occur as special

Table 1. Contents by Sections of Jacobi's book. 192 pages.

Sections	**Topic**
Chap. 1	*Transformation of elliptic functions.*
1, 2	The general transformation problem.
3, 4	The principles of transformation.
5–9	Given an expression $\frac{dy}{\sqrt{\pm(y-\alpha)(y-\beta)(y-\gamma)(y-\delta)}}$, reduce it to the simpler form $\frac{dx}{M\sqrt{(1-x^2)(1-k^2x^2)}}$.
10–12	Transformation of an expression $\frac{dy}{\sqrt{(1-y^2)(1-\lambda^2y^2)}}$ into another similar form $\frac{dx}{M\sqrt{(1-x^2)(1-k^2x^2)}}$.
13, 14	A transformation of order three.
15	A transformation of order five.
16	A transformation applied twice results in a multiplication.
17	A new notation for elliptic functions.
18	Fundamental formulas in the analysis of elliptic functions.
19	Complex values of elliptic functions. The principle of double periodicity.
20	The analytic theory of transformation of elliptic functions.
21–23	Proof of the analytic formulas for the transformation.
24	Different transformations of the same order. Two real transformations, the larger modulus to the smaller and the smaller to the larger.
25	Complementary transformations. How transforming one modulus to another produces a transformation of the complementary moduli.
26, 27	Transformations supplementary to multiplication.
28	General analytic formulas on the multiplication of elliptic functions.
29–34	The effect of the modular equations.
Chap. 2	*The theory of expansion of elliptic functions.*
35–38	The expansion of elliptic functions in infinite products.
39–42	The expansion of elliptic functions in series of sines and cosines of a multiple of the argument.
43–46	General formulas for expanding the functions $\sin^n \mathrm{am}\left(\frac{2Kx}{\pi}\right)$ and $1/\sin^n \mathrm{am}\left(\frac{2Kx}{\pi}\right)$ in series of sines and cosines of multiples of x.
47, 48	Series expansions of the different kinds of elliptic integrals.
49, 50	Indefinite elliptic integrals of third kind reduced to the special case in which the amplitude equals the parameter.
51, 52	Series expansion of the elliptic integral of third kind. How to express the latter conveniently by means of new transcendental Θ functions.
53–55	Addition of the argument, amplitude, and parameter in elliptic integrals of third kind.

Table 1. (*Continued*)

Sections	**Topic**
56–60	Reductions of the expressions $Z(iu)$ and $\Theta(iu)$ to real arguments. Reduction of general elliptic integrals of third kind with complex argument, amplitude, and parameter.
61	Elliptic functions are quotients. The functions H and Θ that form their numerators and denominators.
62–66	Series expansion of H and Θ. A third expansion of the elliptic functions.

cases, and which played so important a role in Jacobi's entry into this field, are noted explicitly. The next two Sections introduce his notation for the inverses of the integrals, after which complex variables are introduced, and the double periodicity of the functions as functions of a complex variable is noted. The theory of transformation is then revisited in this new context, resulting in large numbers of formulas over the next six Sections. The first of the two parts is brought to a close by a discussion of the significance of the differential equations for complementary moduli and the complete integrals.

The second half of the work is devoted to a large variety of representation theorems for elliptic functions. Being analytic, these functions have local Taylor series expansions. Being periodic, they have natural Fourier series. Jacobi began with infinite-product expansions and followed with the Fourier expansions of elliptic functions of first kind, then the analogous expansions for functions of second and third kinds. The most revolutionary and profound part of the entire work, however, was his global representation of these functions as quotients of theta functions. Jacobi introduced these functions as products, and after many preliminary transformations of them finally gave the series representations

$$\begin{aligned}\Theta\left(\frac{2Kx}{\pi}\right) &= 1 - 2q\cos 2x + 2q^4\cos 4x - 2q^9\cos 6x + \cdots;\\ H\left(\frac{2Kx}{\pi}\right) &= 2q^{1/4}\sin x - 2q^{9/4}\sin 3x + 2q^{25/4}\sin 5x - \cdots,\end{aligned} \tag{37}$$

where $q = e^{-\pi K'/K}$. These series are remarkably well adapted to such representations, since they converge very rapidly and have the kind of periodicity that makes their quotient doubly periodic. For example, as Jacobi showed in Section 61,

$$\sin\mathrm{am}\frac{2Kx}{\pi} = \frac{1}{\sqrt{k}}\frac{H(2Kx/\pi)}{\Theta(2Kx/\pi)}. \tag{38}$$

Theta functions were to prove enormously important in the future, right down to the present day, and whole treatises have been devoted to them. Jacobi himself, two decades later, showed how neatly they could be used to express the parameters of the motion of a rigid body free of external torque, and their analogues in several variables were the key tool needed to solve the Jacobi inversion problem, a problem formulated by Jacobi in 1832 on the basis of Abel's general theorem on algebraic integrals. (It was Jacobi who suggested

both the names *Abel's theorem* and the name *Abelian integrals* involved in the Jacobi inversion problem; this was solved in part by A. Göpel and J.G. Rosenhain in the 1840s, and definitively and independently by Bernhard Riemann and Karl Weierstrass around 1855.) The recognition that a representation valid *globally*, for all values of the independent variable, was possible through the use of quotients of theta functions fully justified the phrase 'new foundations' that Jacobi had chosen for his title.

6 SIGNIFICANCE OF THE WORK

Jacobi's treatise represents a clear and thorough exposition of the theory of elliptic functions at the point where it had just crossed the threshold into its proper context, the complex domain. The properties possessed by elliptic integrals and their inverses as functions of a complex variable, especially the property of double periodicity, were now shown to be of great importance; and the inspiring power of theta functions in the study of these functions was clearly established. Still, the full resources of the theory of analytic functions of a complex variable had not yet been applied to them, and consequently elliptic functions were understood very imperfectly. In the five years preceding the publication of the works of Abel and Jacobi in this area Cauchy had been making brilliant advances in the general theory of analytic functions of a complex variable and publishing his results in his *Exercices d'analyse*, which Abel purchased and read eagerly. Cauchy's invention of contour integrals over paths in the complex plane, which occurred in 1825, is a good example (§28).

However, Abel's most far-reaching result had been obtained before he went to Paris. Neither Abel nor Jacobi mentioned Cauchy in connection with elliptic functions. The applicability of Cauchy's work to elliptic functions was not yet clear, and Cauchy hardly ever touched upon it in his many publications. Probably the first real contact between the two topics occurred in 1846, when Cauchy discovered branch points and their connection with periodicity of the inverse function of the integral. Cauchy did not connect the multivaluedness of the integral to that of the integrand (perhaps because the integral is multivalued in an important case where the integrand is not, namely the case of the Cauchy kernel). The 'neglect' of Cauchy by Abel and Jacobi could not have been due to ignorance, since Abel certainly knew of Cauchy's work, and Jacobi, in an article on the roots of polynomials [Jacobi, 1827] praised Cauchy and Fourier for having introduced integrals between arbitrary limits.

This treatise is a 'period piece', marking a stage in the maturation of the theory of elliptic functions. It was a radical departure from the just-completed treatise of Legendre, but it was to be supplanted two decades later by new approaches to the study of algebraic functions in the work of Riemann and Weierstrass. Nevertheless, its importance was recognized for a generation afterward, even after Weierstrass had greatly streamlined the theory. When Jacobi's collected works were published, Carl Borchardt, and then after his death, Weierstrass, took the trouble to republish it, correcting what Weierstrass described as its 'numerous misprints, misstatements, and computational errors'.

Both Riemann and Weierstrass were led to important discoveries by working on the Jacobi inversion problem, which had been stated by Jacobi in connection with his analysis of Abel's theorem, and the solution of the inversion problem had been achieved only by

generalizing theta functions to several variables. The work of Abel and Riemann showed that algebraic curves of the same genus can be treated in a unified manner. In particular, one could move one of the roots of a fourth-degree polynomial to infinity and replace that polynomial with a cubic. Such an approach allowed Weierstrass to derive the important properties of elliptic functions from the mere fact of their double periodicity. Even after the more sophisticated approaches of Riemann and Weierstrass came to be applied, however, Jacobi's treatise remained a valuable source of the concrete examples on which the more general theories were modeled. The theta functions that Jacobi introduced proved to be an extremely fertile source of information, and have been generalized greatly in the 20th century. What might appear to be a more modest innovation, the Jacobi elliptic functions, are still important in applications. In his two-volume treatise on analytic function theory, the late Einar Hille devoted a long chapter to elliptic functions, remarking [1959, vol. 2, 144], that

> So far the functions of Weierstrass have been kept in the foreground, but neither are they the oldest elliptic functions discovered, nor are they the ones best suited for applications to arithmetic, geometry, and mechanics. The functions introduced by Jacobi antedate those of Weierstrass by some thirty years, and the student cannot afford to ignore them.

Elliptic functions have found a niche in the general theory of algebraic functions, one of the most important in applications, as the recent proof of Fermat's Last Theorem amply demonstrates. Perhaps the fact that they are a special case of the general theory accounts for the neglect of this topic in the extensive 450-page report [Brill and Noether, 1893] on the development of algebraic functions rendered to the *Deutsche Mathematiker-Vereinigung*. In that report Legendre was mentioned only once and only in passing. His *Traité* was not mentioned at all, nor was Jacobi's *Fundamenta*. On the other hand, a very full account is given in Volume 2 of the *Encyclopädie der mathematischen Wissenschaften*, published two decades later [Fricke, 1913].

The person best able to judge Jacobi's work was the aged Legendre, who had spent a good portion of his life trying to bring some order into the theory of elliptic functions. In his time the subject was in its infancy, and like all infantile scientific theories, the theory of elliptic functions was most in need of classification and taxonomy. These Legendre had provided, showing how to reduce any elliptic integral to a sum of three standard kinds. The reason why just these three kinds occur could not be explained fully until the time of Riemann and Weierstrass; but the fact of their existence provided a framework in which further work could be done. (Further, when it was done, two of the resulting three kinds of integrals were not exactly what Legendre had defined: see Fricke [1913, 189].) Legendre was extremely generous in his praise of the work of Abel and Jacobi; he was indeed very glad to see that his 40 years' work on this subject, which effectively created the subject of elliptic functions as an area for research, would not be forgotten. All this he expressed both at the Paris Academy and to other scholars, including Alexander von Humboldt, who wrote to Jacobi on 2 April 1828 [Pieper, 1987, 47]:

> I am most profoundly happy for the great respect you have attained among the leading mathematicians of our time through your excellent work. I have sent

> Herr Le Gendre's letter, which is full of lavish praise for your discovery, to the Minister of Culture, and I plan to present a copy of it to the Academy of Sciences this week. For me it is a pleasant duty to give you this proof of my liveliest interest in your work.

What would have become of the Jacobi–Abel rivalry had Abel lived is impossible to know. What is certain is that after Abel's death Jacobi worked diligently to assure that all due credit should be given to Abel for his epoch-making paper on general algebraic integrals. In his memorial tribute to Jacobi in 1852, Dirichlet commented thus [Jacobi, *Works*, vol. 1, 13],

> Given that Abel and Jacobi improved the theory simultaneously in two different directions, it appears that Fate wished to divide the honor of the advances to be made equally between the two young rivals; for the manner in which each soon extended the discovery of the other leaves no doubt that either of them would have achieved the entire advance alone.

BIBLIOGRAPHY

Abel, N.H. 1827. 'Recherches sur les fonctions elliptiques', *Journal für die reine und angewandte Mathematik*, *2*, 101–181. [Repr. in *Oeuvres complètes*, 2nd ed. (1881), vol. 1, 141–221.]

Brill, A. and Noether, M. 1893. 'Die Entwicklung der Theorie der algebraischen Functionen in älterer und neuerer Zeit', *Jahresbericht der Deutschen Mathematiker-Vereinigung*, *3*, 109–566.

Cooke, R. 1989. 'Abel's theorem', in D.E. Rowe and J. McCleary (eds.), *The history of modern mathematics*, Boston: Academic Press, 389–424.

Euler, L. 1761. 'De integratione aequationis differentialis $\frac{m\,dx}{\sqrt{1-x^4}} = \frac{n\,dy}{\sqrt{1-y^4}}$', *Novi Commentarii Academiae Petropolitanae*, *6* (1756–1757), 37–47. [Repr. in *Opera omnia*, ser. 1, vol. 20, 58–79.]

Fricke, R. 1913. 'Elliptische Funktionen', in *Encyclopädie der mathematischen Wissenschaften*, vol. 2, pt. 2, Leipzig: Teubner, 177–349 (article IIB3).

Hille, E. 1962. *Analytic function theory*, vol. 2, Boston: Ginn and Company. [Repr. New York: Chelsea, 1973.]

Houzel, C. 1978. 'Fonctions elliptiques et intégrales abéliennes', in J. Dieudonné (ed.), *Abrégé d'histoire des mathématiques 1700–1900*, Paris: Hermann, vol. 2, 1–113.

Jacobi, C. *Works. Gesammelte Werke*, 7 vols., Berlin: Reimer, 1881–1891. [Repr. New York: Chelsea, 1969.]

Jacobi, C. 1827. 'Ueber die Ausdruck der verschiedenen Wurzeln einer Gleichung durch bestimmte Integrale', *Journal für die reine und angewandte Mathematik*, *2*, 1–8. [Repr. in *Works*, vol. 6, 12–20.]

Klein, F. 1926. *Vorlesungen über die Entwicklung der Mathematik im 19. Jahrhundert*, vol. 1, Berlin: Springer-Verlag. [Repr. New York: Chelsea, 1969.]

Ore, O. 1957. *Niels Henrik Abel, mathematician extraordinary*, Minneapolis: University of Minnesota Press. [Repr. New York: Chelsea, 1974.]

Pieper, H. (ed.) 1987. *Briefwechsel zwischen Alexander von Humboldt und C.J. Jacob Jacobi*, Berlin: Akademie-Verlag.

CHAPTER 32

HERMANN G. GRASSMANN, *AUSDEHNUNGSLEHRE*, FIRST EDITION (1844)

Albert C. Lewis

This book contained the basis for what came to form vector analysis and linear algebra. Though its direct influence on these subjects developed slowly, it played a seminal role in some key works by others. Grassmann died in 1877 just as his work was gaining recognition.

First publication. *Die lineale Ausdehnungslehre, ein neuer Zweig der Mathematik, dargestellt und durch Anwendungen auf die übrigen Zweige der Mathematik, wie auch auf die Statik, Mechanik, die Lehre vom Magnetismus und die Krystallonomie erläutert,* Leipzig: Otto Wigand, 1844. xxxii + 279 pages.

Manuscript. Versions are described in [Grassmann, *Works*], but the *Nachlass* is no longer extant.

New edition. As *Hermann Grassmanns gesammelte mathematische und physikalische Werke*, vol. 1, part 1 (ed. F. Engel), Leipzig: Teubner, 1894–1911. [Photorepr. New York: Chelsea, 1969; New York and London: Johnson Reprint, 1972.]

French translation. *La science de la grandeur extensive: la 'Lineale Ausdehnungslehre'* (trans. D. Flament and B. Bekemeier), Paris: Blanchard, 1994.

English translation. *A new branch of mathematics: the* Ausdehnungslehre *of 1844 and other works* (trans. L.C. Kannenberg), Chicago and La Salle, Illinois: Open Court, 1995.

Spanish translation. Teoría de la Extensión. Nueva disciplina matemática expuesta y aclarada mediante aplicaciones (trans. E. Oscar Roxin), Buenos Aires: Espasa-Calpe Argentina, 1947.

Related articles: Hamilton (§35), Thomson and Tait (§40).

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

1 A FAMILY OF MATHEMATICIANS

The principal influence in the mathematical education of Hermann Grassmann (1809–1877) was his father, Justus Günther (1779–1852), who was a Gymnasium teacher in Szczecin (Stettin), Pomerania (now a part of Poland). He attended university to study theology and philology but on the way to obtaining qualification as a teacher turned to mathematics as his life's main work. Of his twelve siblings, Grassmann appears to have worked most closely with his younger brother Robert and the two may have collaborated to some degree on foundational and philosophical aspects of the *Ausdehnungslehre*; however, it appears that they went somewhat separate ways after 1844, Robert going on to pursue logic and the philosophy of science.

Like his father, Grassmann remained a Gymnasium teacher for the whole of his career. Though he sought university positions, this was mainly for their academic environment; with respect to compensation and prestige, his school-teaching position provided at least as well as a university position would have. In addition to the mathematical work following upon the *Ausdehnungslehre* that is described below, Grassmann made significant contributions to physics and philology. He published textbooks in mathematics, German and Latin, the most important mathematically being the *Lehrbuch der Arithmetik* in 1861 whose general logical structure and proofs were utilized by Giuseppe Peano in his work on the foundations of the number system in 1889 (§47.4). In between Grassmann produced a major dictionary on Sanskrit. Though less available and not intended for a wide audience, his school programs (series of formal treatises by teachers intended as pedagogical resources) on these same topics and others reveal aspects of the philosophical approach that underlay much of his more widely known work.

The main source of information on Grassmann's life and work is [Grassmann, *Works*]; the present account draws upon the biography in vol. III, part 2, if no other source is given. A comprehensive introduction to Grassmann and the *Ausdehnungslehre* in particular is provided by the proceedings of the conference held on the island of Rügen, not far from Szczecin, on the occasion of the sesquicentenary of the *Ausdehnungslehre* [Schubring, 1996a]. The place of the *Ausdehnungslehre* in the development of vector analysis, including an account of competition between vectors (Grassmann, J.W. Gibbs) and quaternions (W.R. Hamilton and followers) is given in [Crowe, 1967]; see also (§35).

2 DIALECTICS AND THE THEORY OF TIDES

One of the major influences on Grassmann by his own account was the theologian F. Schleiermacher, one of his professors at Berlin University. Hermann and his brother Robert read Schleiermacher's *Dialektik* in 1840, and there are signs of the influence of that work in the *Ausdehnungslehre*. It is possible that, through Robert, Hermann was also influenced by the German philosopher J.F. Fries [Schubring, 1996b]. Grassmann's first mathematical work, that decisively launched him into a career in mathematics, was a treatise on the theory of tides [Grassmann, 1840]. Though it was composed in 1840 as a part of his examination for entry into the teaching profession, it was not published until the *Werke* edition in 1911 when its relevance as a precursor to the *Ausdehnungslehre* became

evident. In addition to exhibiting an understanding of the theory of tides as based on the work of J.L. Lagrange and P.S. Laplace, Grassmann presented a new way of casting the mathematics involved. To take a simple example, the acceleration at a point represented by the coordinates x, y, and z, for example, were traditionally represented by three equations expressing the three components of acceleration:

$$\frac{d^2}{t^2}x = X + X_1 + X_2 + \cdots, \quad \frac{d^2}{t^2}y = Y + Y_1 + Y_2 + \cdots, \quad \frac{d^2}{t^2}z = Z + Z_1 + Z_2 + \cdots. \quad (1)$$

Grassmann introduced what has since been called a vector notation, which reduced these to the single equation:

$$\frac{d^2}{t^2}p \doteq P \dot{+} Q \dot{+} R \dot{+} \cdots. \quad (2)$$

Conceptually for Grassmann this involved considerably more than simply a change of notation: the view is shifted from purely numerical relationships to geometrical ones. Though in some respects this offered a greater simplicity, it raised the issue of recognizing different types of equality and types of addition and multiplication (signaled by the dots above the operators in this last equation) depending on the different types of entities (points, lines and areas for example), and the operations between various combinations of these entities. This challenging complication was very probably a factor that led to the greater abstraction of the *Ausdehnungslehre*, in particular to the universal algebra aspect of its 'theory of forms' [Lewis, 1996a].

In the *Ausdehnungslehre* Grassmann attempted to provide a foundation that involved also a style of presentation unusual in the tradition of mathematical works in that it tried to reflect his actual path of discovery rather than just a formal presentation of the results of discovery. His 'foundation' was more philosophical than logical from the point of view of the direction the subject took since his time. Apparently bolstered by Schleiermacher's dialecticism, Grassmann held that the identity and unity of mathematics came from a system of contrasts at all levels: from mathematics vis-à-vis philosophy and the other sciences, to the presentation of the different types of multiplication [Lewis, 1977]. For example, he posits four principal branches of mathematics that correspond to contrasting types of elements and modes of generation of those elements. Grassmann's exposition of this classification, in which he places the calculus of extension, is summarized in Table 1.

Table 1. Grassmann's classification.

		Mode of generation	
		positing and connecting	single generation
Type of element	equal	algebraic-discrete (number theory, arithmetic)	algebraic-continuous (functions, calculus)
	different	combinatorial-discrete (combinatorial analysis)	combinatorial-continuous (calculus of extension)

3 THE NEW BRANCH OF MATHEMATICS

The *Ausdehnungslehre* was the first Part of a planned larger work announced on the general title page: *Die Wissenschaft der extensiven Grösse oder die Ausdehnungslehre, eine neue mathematische Disciplin dargestellt und durch Anwendungen erläutert von Hermann Grassmann*. The second Part was to address non-linear aspects, though it is not very clear what they might have been as it never appeared. The contents of the first Part are summarized in Table 2.

It is not easy to do justice in a translation to Grassmann's original vocabulary. As a linguist he was attentive to the choice of terms chosen for the new concepts and expressed a strong preference for using words that shared etymological roots with everyday German language. A translation could adopt a modern usage to make the text as accessible as possible to a modern mathematical reader or it could tend towards using unusual terms by today's standards that would convey something of the novelty Grassmann's readers presumably encountered. Thus, in English Grassmann's '*Strecke*' might be rendered 'directed line segment', 'displacement', 'sect', even 'vector'; however 'stroke' or 'stretch' might be more faithful to his stylistic intentions. The following sketch of some of the topics raised in the *Ausdehnungslehre* attempts to convey his style while also giving an indication in modern language of several results that can be immediately recognized as part of the standard literature today. The informed reader may see, even in this very selective synopsis, key

Table 2. Contents by chapter of Grassmann's book.

Chapter	**Page**	**Contents**
Foreword	v	Grassmann's first steps towards a new branch of mathematics.
Introduction	xix–xxxii, 1	The conception of the *Ausdehnungslehre* as a science; the general theory of forms.
I.1	15	Addition and subtraction of displacements.
I.2	47	Outer multiplication of displacements.
I.3	74	Connection of extensive magnitudes of higher step.
I.4	90	Outer division and number magnitude.
I.5	114	Equalities and projections.
II.1	130	Addition and subtraction of elementary magnitudes of first step.
II.2	147	Outer multiplication, division and shadow of elementary magnitudes.
II.3	182	Regressive product.
II.4	229	Relations between various relations (e.g., shadow and projection; the shadow of a product).
Note	266	On the open product.
Table of contents	275–279	

results of linear algebra and the unfolding of what has since been named the Grassmann or exterior algebra.

Grassmann made a distinction between the subject of geometry, as conceived in the early 19th century as the science of space, and a purely mathematical, abstract foundation for that geometry. His *Ausdehnungslehre*, or calculus of extension, was to be that foundation, and it might be noted that geometry is otherwise not represented in the above classification table. This new branch has an algebraic or combinatorial aspect, though it is capable of dealing with continuous geometrical entities. Unlike ordinary algebra it is capable of representing geometrical dimension and, as an abstract theory, is not confined to the three dimensions of physical space [Lewis, 1997]. For example, the general result is obtained that if a field ('*Gebilde*') or system of order ('*Stufe*') a and one of order b are contained in a field of order c, but in no field of order less than c, then the first two fields will intersect in a field of order at least $a + b - c$. In three-dimensional geometry, where the point, line, plane, and solid are fields of order 1, 2, 3, and 4 respectively, this result is exemplified by their various intersection possibilities.

The *Ausdehnungslehre* also combines the two approaches to geometry: synthetic and analytic. As an example of this, consider the vertex γ of a triangle whose other vertices α and β move in fixed lines A and B, and whose sides opposite α, β and γ pass through three fixed points a, b, and c respectively. It is known that the vertex γ describes a conic. In the *Ausdehnungslehre* the equation of the conic can be written as follows where, it should be noted, the expression is of degree two in γ:

$$\gamma a B c A b \gamma = 0. \tag{3}$$

The product of two points is the line joining them, and the product of two lines is their point of intersection. Hence the product $\gamma a B$ represents the vertex β while $\gamma a B c A$ represents the vertex α. The product of three points is the triangle formed by them and thus when the three are collinear, as in this case for α, b, and γ, their product is zero (art. 147).

The simplest rules underlying these geometrical examples are established from general rules of connection or operation ('*Verknüpfung*') in the following way. A connection of first order, including what Grassmann termed its synthetic and analytic forms, follows the usual arithmetic rules of commutativity and associativity. The next higher order of connection, is determined by the property of distributivity to the next lower order. Thus if $\cap$ represents the first order connection and $\Cap$ the second order connection, the relationship between the two is expressed by Grassmann as:

$$(a \cap b) \Cap c = (a \Cap c) \cap (b \Cap c). \tag{4}$$

The particular sorts of addition and multiplication (as well as their converses subtraction and division) used will be written hereafter in the *Ausdehnungslehre* in the usual arithmetic way. Multiplication, however, is associative but not necessarily commutative, only right and left distributive over addition.

Displacements ('*Strecken*') are extensives of the first order and a system of first order is generated by addition and subtraction of a displacement or parts of a displacement. Displacements can be generated independently of each other, that is in such a way that one

cannot be expressed as a sum of the others (art. 19). Systems of higher order are constructed by addition of mutually independent displacements. If

$$p_1 = a_1 + b_1 + \cdots \quad \text{and} \quad p_2 = a_2 + b_2 + \cdots \tag{5}$$

are two displacements expressed as sums of independent displacements, then their addition is represented by

$$p_1 + p_2 = (a_1 + a_2) + (b_1 + b_2) + \cdots. \tag{6}$$

Grassmann demonstrates that a system of order m can be generated from any collection of m independent displacements, and in the course of this he presents what is now recognized as one of the main theorems dealing with dimension of a vector space: if a system of order m is generated by m independent extensives then the system cannot contain a system of independent extensives of order greater than m (art. 20).

Chapter One concludes with examples of applications in mechanics along the lines of those outlined in his earlier work on the theory of tides. Here Grassmann reiterates that brevity is not the most important advantage of his new subject, rather it is that each calculational step is a 'pure expression of the corresponding conceptual step' (art. 27).

Multiplication is introduced in the next chapter and motivated by considering a linear displacement being shifted along another in the same fixed plane and thereby sweeping out a parallelogram. Two such areas will have the same sign 'if, in passing from the direction of the moved displacement to the direction of that constructed by the motion, both inflect to the same side (for example, to the left), but opposite signs if they are oppositely inflected' (art. 28). Considered as a connection between the moved displacement and fixed displacement, this notion is shown to have the distributive properties described above and thus to qualify as multiplication. With this as a concrete example, Grassmann proceeds to formally develop the general notion of 'outer multiplication' which, in addition to the left and right distributive properties and associativity, have the properties that $ab = -ba$ and $aa = 0$ for extensives a and b. One of the applications making use of this multiplication is to solve systems of linear equations (arts. 45, 46).

Chapter Three extends the notion of addition. Up to this point the extensives being added were of the same order whether displacements (first order) or products of displacements (higher orders). Here Grassmann develops what it means to add extensives of different orders and makes the definition hinge on their common factors of lower order. He regards his most dramatic application of the ideas of this chapter to be in moment problems in mechanics where the collective moment of several forces with respect to a point can be represented as the sum of all the individual moments with respect to that point.

The analytic connection corresponding to outer multiplication is outer division. In Chapter Four the general procedure for defining division is followed, namely as the determination of one factor in terms of the product and the other factor. Here there are pitfalls in looking to arithmetic for too much guidance even if the results in the end appear closely analogous. The main difference is that the numerical magnitudes for Grassmann are developed as quotients of extensives rather than built up from discrete numbers as arithmetic might be built up from the integers.

The last chapter of Part I is concerned with solving equations involving the entities developed thus far. It introduces the process of projection ('*Projektion*') or its more general version, shadow ('*Abschattung*').

Part II deals with 'elementary magnitudes' and resembles in some aspects A.F. Möbius's barycentric calculus. Though developed independently of Möbius and in a more general way, it too can be interpreted geometrically as a system consisting of weighted points in space and operations on them. The elementary magnitude of weight zero is shown to be interpretable as a displacement. The operations from Part I are given a meaning for these elementary magnitudes and a new product is introduced. Grassmann first reminds us that the outer product was essentially characterized as having a non-zero result when the factors were mutually independent (i.e. 'outside' of each other) and zero if they were dependent. The new, regressive ('*eingewandte*') product allows for the possibility of a non-zero result for dependent factors. Since higher-order factors can have lower-order common components there is actually a range of regressive products corresponding to the range of possible orders. The key to establishing rules for this multiplication is the relation between the orders of the step common to the factors and the order of the nearest step covering the factors. Here Grassmann gives the relationship which, in modern linear algebra, would express the relationship between dimensions of subspaces: the dimension of the sum of two subspaces plus the dimension of their intersection is equal to the sum of their dimensions (art. 126).

Grassmann's definition of division with respect to a 'system' considers that if $BC = A$ then C, written as the expression A/B, can be regarded as a quotient but only with the understanding that the 'complete' quotient involves a component which is dependent on B (art. 141). This complete quotient, which he writes as $A/B + 0/B$, has been taken by later commentators as an expression of abstraction (i.e. an equivalence class). John Venn noted in his *Symbolic logic* that this notion anticipated almost exactly the same form of expression in Boole ([Venn, 1881, 204]; compare §36.7).

The final section of the work is entitled 'Note on open products' (art. 172) and contains what J.W. Gibbs referred to as 'the key to the theory of matrices' [Gibbs, 1891, 81]. Gibbs argued that the later work of Cayley and Sylvester in developing an algebra of matrices was anticipated, and even treated in a more general fashion, in this 'note' to the *Ausdehnungslehre*.

4 A MUTED RECEPTION

Grassmann sent copies of his new publication to several mathematicians including the Frenchman B. de Saint-Venant (whose work Grassmann had read), and the most renowned mathematician of the time, C.F. Gauss. Not knowing Saint-Venant's address, Grassmann sent the book through A.L. Cauchy who seems never to have sent it on; at least Saint-Venant later informed Grassmann that he had never received it. It was of some concern to several mathematicians acquainted with the *Ausdehnungslehre* that shortly afterwards Cauchy presented his 'clefs algébriques', which bore a striking resemblance to parts of Grassmann's work. Though the issue of possible plagiarism was brought before the French Academy of Sciences, there was no resolution of the question. Gauss, on the other hand,

courteously replied that he had skimmed through the book but because of the press of other work he could not find the time that evidently would be required to fully understand it. Grassmann's friend Möbius, whose work was so closely related to his program, had to admit that he found the *Ausdehnungslehre* too philosophical. No reviews or notices of the work were published.

Möbius did, however, encourage Grassmann to submit something for the Jablonowski Science Society prize, which was offered in 1845 for a work that would help demonstrate, in effect, the feasibility of the geometrical calculus envisioned by Leibniz whereby geometrical objects and their relationships were represented by simple symbols without reference to the magnitudes of lines or angles. Grassmann's *Geometrische Analyse* (published with an introduction by Möbius in 1847; reprinted in *Works*, vol. I, part 1) won the prize, and the fact that he was the only contender did not detract from the boost that this gave to the *Ausdehnungslehre* upon which his prize work drew. It was on the basis of this work, for example, that Luigi Cremona in Italy publicized Grassmann's work in 1860.

The only mathematician whose work came close to competing with Grassmann's was that of the Irishman W.R. Hamilton. As the discoverer of quaternions in 1843 (§35) he was an attentive reader when he came across the *Ausdehnungslehre* in 1853. Hamilton's annotated copy is in the Graves Collection in the library of University College London, bound with Hamilton's copy of Möbius's *Der barycentrische Calcul* (1827).

Copies of the 1844 edition are rather scarce since the publisher shredded their stock of unsold copies fairly soon. It should be kept in mind that the means for communicating mathematical research in Europe in the 1840s were limited compared with the situation a few decades later when journals and societies began to proliferate.

5 EVENTUAL RECOGNITION

On the basis of numbers of publications Grassmann could appear to have turned away from the *Ausdehnungslehre* after its publication and towards philology, and Sanskrit studies in particular. Though this might be understandable in view of the poor reception given the *Ausdehnungslehre*, he actually managed to revise its presentation and demonstrate its applicability in a wide range of subjects. In his search for such applications in the 1850s he made some significant contributions to colorimetry. In 1861 he published a completely rewritten version that dispensed with the philosophical, pedagogical underpinnings that he tried to integrate into the original presentation. (This version is often listed as 1862, the title page date, but it is known that Möbius received a copy in October 1861. It is reprinted in *Works*, vol. I, part 2.) In particular, he utilized a Euclidean style of presentation, precisely what he had tried to avoid in 1844. This streamlined version provided an introduction to the *Ausdehnungslehre* for many who never looked at or even knew about the first version. Hermann Hankel made extensive and admiring use of the 1861 version in his influential *Theorie der complexen Zahlensysteme* (1867). From Hankel, Felix Klein came to learn of Grassmann, and Klein in turn brought Grassmann to the attention of Alfred Clebsch. He and, especially after his death in 1872, his students were a significant force in propagating the ideas of the *Ausdehnungslehre*. Among them, in addition to Klein, were Paul Gordan, Alexander Brill, Olaus Henrici, Max Noether, and Ferdinand Lindemann [Tobies, 1996].

Thus Grassmann, in the last years of his life, saw the *Ausdehnungslehre* being recognized. He revived the original version in a new edition that appeared posthumously in 1878.

After Grassmann died in 1877 his work gradually became known beyond Germany. The English mathematician W.K. Clifford predicted in 1878 that the *Ausdehnungslehre* 'will exercise a vast influence upon the future of mathematical science' [Clifford, 1878]. The American philosopher and mathematician C.S. Peirce published a note relating Grassmann's external and internal products to quaternions and to Peirce's own algebras [Peirce, 1877–1878]. Paul Carus, founding editor of the Open Court Press in the United States, was a student of Grassmann in Szczecin, and claimed inspiration from Grassmann's mathematical work in his own work on the unity of science and religion [Carus, 1889]. In 1898 the Englishman Alfred North Whitehead based his *A treatise on universal algebra* largely on applications of Grassmann's calculus of extension to various geometries. Thanks to these and other seminal writings, by the end of the 19th century the *Ausdehnungslehre* was generally recognized as one of the principal mathematical treatises in the history of mathematics.

At the instigation of Klein, Grassmann's mathematical and physical works were collected together in the *Werke*, a model of a scholarly edition. Even if the principal editor, F. Engel, was not particularly sympathetic to Grassmann's full program, he enlisted the help of the most qualified 'Grassmannians' of the time in producing a worthy monument and resource. Its annotated edition of the 1844 *Ausdehnungslehre*, for example, provides a collation with the 1878 version.

Much of the *Ausdehnungslehre* of 1844, especially if seen from the standpoint of the 1861 rewriting, contains abstract notions, such as a not necessarily commutative ring, that are generally thought of as 20th-century constructions. Gian-Carlo Rota enjoyed telling how the 'discoveries' of his and some other modern mathematicians working in combinatorics were, as he determined by his own reading of the 1861 version of the *Ausdehnungslehre*, actually rediscoveries [Stewart, 1986]. One commentator has remarked that practically all Grassmann lacked relative to the more modern developments was the language of sets [Fearnley-Sander, 1979]. In addition to so many ideas ahead of their time, the formal aspect of Grassmann's approach, relying as it does on the implicit definition of mathematical entities, can be viewed as a move in the direction of modern axiomatics. But, as indicated above, it seems that Grassmann himself would likely not have favored characterizing the nature of mathematics solely in terms of axiomatic developments. Like many other classical works in the history of science, the 1844 *Ausdehnungslehre* carries within itself a separate world with a life of its own.

BIBLIOGRAPHY

Carus, P. 1889. 'The old and the new mathematics', *Open court, 2*, 1468–1472.

Clifford, W.K. 1878. 'Applications of Grassmann's extensive algebra', *American journal of mathematics, 1*, 350–358.

Crowe, M.J. 1967. *A history of vector analysis: The evolution of the idea of a vectorial system*, Notre Dame: University of Notre Dame Press. [Repr. New York: Dover, 1985.]

Dorier, J.-L. 1996. 'Basis and dimension—from Grassmann to van der Waerden', in [Schubring, 1996a], 175–196.

Fearnley-Sander, D. 1979. 'Hermann Grassmann and the creation of linear algebra', *American mathematical monthly, 86*, 809–817.

Gibbs, J.W. 1891. 'Quaternions and the "Ausdehnungslehre" ', *Nature, 44*, 79–82. [Repr. in *The scientific papers*, vol. 2, 161–168.]

Grassmann, H.G. *Works. Gesammelte mathematische und physikalische Werke* (ed. F. Engel et alii), 3 vols., each in 2 pts., Leipzig: Teubner, 1894–1911. [Repr. New York and London: Johnson Reprint, 1972.]

Grassmann, H.G. 1840. *Theorie der Ebbe und Flut, Prüfungsarbeit.* [Repr. in Grassmann *Works*, vol. III, part 1.]

Lewis, A.C. 1977. 'H. Grassmann's 1844 *Ausdehnungslehre* and Schleiermacher's *Dialektik', Annals of science, 34*, 103–162.

Lewis, A.C. 1981. 'Justus Grassmann's school programs as mathematical antecedents of Hermann Grassmann's 1844 *Ausdehnungslehre'*, in H.N. Jahnke and M. Otte (eds.), *Epistemological and social problems of the sciences in the early nineteenth century*, Dordrecht: Reidel, 255–268.

Lewis, A.C. 1996a. 'The influence of Grassmann's 1840 theory of tides on the *Ausdehnungslehre'*, in [Schubring, 1996a], 29–36.

Lewis, A.C. 1996b. 'Some influences of Hermann Grassmann's program on modern logic', in I. Angelelli and M. Cerezo (eds.), *Studies on the history of logic: Proceedings of the III symposium on the history of logic*, Berlin and New York: de Gruyter, 377–382.

Lewis, A.C. 1997. 'Hermann Grassmann's n-dimensional vector concept', in D. Flament (ed.), *Le nombre, une hydre à* n *visages: entre nombres complexs et vecteurs*, Paris: Maison des Sciences de l'Homme, 139–148.

Peirce, C.S. 1877–1878. 'Note on Grassmann's calculus of extension', *Proceedings of the American Academy of Arts and Sciences, 13*, 115–116.

Schubring, G. (ed.) 1996a. *Hermann Günther Grassmann (1809–1877): Visionary mathematician, scientist and neohumanist scholar. Papers from a sesquicentennial conference*, Dordrecht: Kluwer.

Schubring, G. 1996b. 'The cooperation between Hermann and Robert Grassmann on the foundations of mathematics', in [Schubring, 1996a], 59–70.

Stewart, I. 1986. 'Hermann Grassmann was right', *Nature, 321*, 17.

Tobies, R. 1996. 'The reception of H. Grassmann's mathematical achievements by A. Clebsch and his school', in [Schubring, 1996a], 117–130.

Venn, J. 1881. *Symbolic logic*, London: Macmillan.

CHAPTER 33

KARL GEORG CHRISTIAN VON STAUDT, BOOK ON PROJECTIVE GEOMETRY (1847)

Karin Reich

In this book Staudt tried to 'purify' the principles of projective geometry by removing all metrical notions. Thereby he also raised synthetic geometry to a new level. He laid emphasis on involution, with his influential quadrilateral construction. Together with Poncelet, Gergonne and Steiner, he belongs to the founders of projective and synthetic geometry.

First publication. Geometrie der Lage, Nürnberg: Verlag von Bauer und Raspe (Julius Merz), 1847. Also Nürnberg: Verlag Fr. Korn, [no date]. vi + 216 pages.

Italian translation. *Geometria di posizione* (trans. M. Pieri), Turin: Fratelli Bocca, 1889.

Related articles: Monge (§17), Poncelet (§27), Klein (§42), Hilbert on geometry (§55).

1 BACKGROUND AND BIOGRAPHY

Building upon some aspects of the descriptive geometry of Gaspard Monge (1746–1818) and his followers, Jean Victor Poncelet (1788–1867) developed a new view of geometry in his *Traité des propriétés projectives des figures* of 1822, with his emphasis upon poles, polars, reciprocal polars, duality and classes of curves (§27). While Poncelet thought that he was the first who had recognized the importance of duality, Joseph Diaz Gergonne (1771–1859) claimed priority. The main contribution of Jakob Steiner (1796–1863) was the projective generation of the conic sections, in his book [Steiner, 1832]. None of these geometers was able (or maybe willing) to present a consequent development of projective geometry, nor were their theories free from metrical considerations. It was von Staudt was the first who adopted a fully non-metrical approach.

Karl Georg Christian von Staudt was born on 24 January 1798 in Rothenburg ob der Tauber, south of Würzburg in southern Germany. He studied at the University of Göttingen, matriculating on 5 May 1819, when Carl Friedrich Gauss (1777–1855), professor of astronomy and director of the observatory, became his main teacher. At first mainly interested in astronomy, von Staudt received a doctorate degree from the University of Erlangen

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

in this field in 1822. In the same year he qualified as a mathematics teacher at the University of Munich. At first he taught at a secondary school in Würzburg, but he also gave lectures at the University. During this time he published an essay [Von Staudt, 1825] on the sectioning of the circle following the method of Gauss. In 1827 Staudt moved to Nürnberg where he gave lectures at a secondary school and at the polytechnical school. He also published an article [Von Staudt, 1831] on curves of the second order, which was his first step in the direction of his later researches. In 1835 he was appointed professor of mathematics at the University of Erlangen, a position that he held until his death on 1 June 1867. On his life see especially [Böhmer, 1953].

2 VON STAUDT'S 'GEOMETRY OF POSITION'

2.1 A tough textbook. The book under notice here was von Staudt's initial publication in projective and synthetic geometry. It is not clear why it came out from two different Nürnberg publishers, one with an undated title page. The texts of the two printings are identical, and the author dated his preface to August 1847.

Although in that preface von Staudt conveyed the impression that he was publishing a textbook, he rendered his theory in a very strict form, so it is hard to read and understand. There are no illustrations, applications, sketches, or references to other work, nor is a foundation laid down on axioms; and many new technical terms are used, not always with explanation. His title, 'Geometry of position', surely alludes to that of Lazare Carnot's pioneering volume *Géométrie de position* of 1803. The contents of the book are summarised in Table 1; on its historical context and initial influence, see especially [Kötter, 1898, pts. 2–3].

Von Staudt was convinced that his geometry of position was a pure projective geometry, more fundamental than other forms of geometry in being free from metrical considerations or of measurement, especially distance and congruence, and notions dependent upon them, such as cross-ratio. His work has a topological feel, as its title implies, although he left intuitive the underlying assumptions (compare §76).

2.2 Preliminary definitions. In chs. 1–4 von Staudt introduced the elements of his geometry, that is, angles, points, lines, surfaces and planes, and solids. He discussed 'bundles of straight lines' ('*Strahlenbündel*'), co-punctual collections of straight lines in space; a planar section of such a bundle was a 'bushel' of straight lines (art. 22). Among other notions, a solid angle was part of a half-bundle (art. 11), and a bushel of planes was specified by the property of sharing a common straight line (art. 23).

Ch. 5 was devoted to vanishing elements such as vanishing points, the vanishing line and the vanishing plane. This means that in the case of two straight lines, placed on the same plane, either they intersect in a point; or they are parallel lines, that is, they have a common direction and intersect in an infinitely distant 'vanishing' point. Von Staudt assumed that the locus of all vanishing points in a plane was a vanishing straight line; all vanishing points and lines were located on an infinitely distant plane (arts. 54–57). He also worked with parallel bundles of straight lines (art. 42). His theory assumed Euclid's parallel axiom (art. 31).

Table 1. Contents by chapters of von Staudt's book.

Ch.	Page	Art.	Topics
	iii–vi		Preface, contents.
1–2	1	1	'Bundles' and 'bushels' of straight lines. Plane-bushels. Space- and surface-angles.
3	13	31	'On parallels'.
4	18	43	'On n corners, *n* edges and polyhedra'.
5	23	54	'Infinitely distant elements'.
6	30	66	'Law of reciprocity'.
7	36	79	'On n corners, *n* edges etc. in another denotation'
8	43	93	'Harmonic configurations'.
9–10	49	103	'Projective relationships between uniform configurations'.
10	60	121	'Projective relationships between fundamental configurations of the second level, and between spatial systems'.
11	72	139	'On lines, surfaces, and configurations related to them'.
12	81	153	'Division of closed lines, surfaces, etc. into such of even and into such of odd order'.
13–14	90	169	'On plane figures and configurations related to them'; and for bodies.
15–17	110	197	'Reversion elements', 'Involution' and 'Involutory systems'.
18	131	234	'Polar systems in the plane and in the straight-line bundle'.
19	137	246	'Curves and conic surfaces of the 2nd order'.
20	149	264	'Projective relationships between curves of the 2nd order'.
21	165	284	'On the number of common points and tangents of two curves of the 2nd order'.
22–23	172	296	'On lines of the 2nd order in general'; exercises.
24	190	318	'Polar systems in space'.
25	197	328	'Surfaces of the 2nd order'.
App.	203	336	Similarity and affinity between figures. [End 216, art. 360.]

2.3 The 'principle of duality'. Von Staudt called it 'law of reciprocity' (ch. 6); it is a main idea in projective geometry. According to him (and his predecessors) points and planes are dual terms; in any theorem one can interchange the words 'point' and 'plane', and also 'connection' and 'intersection', and obtain a new theorem. Poncelet initiated the practise of writing dual theorems in two parallel columns on the page, and Steiner also adopted it. In his foreword von Staudt pointed out that the law of reciprocity is a very useful and attractive feature for pupils in understanding geometry; so he too presented many theorems and exercises this way, for example (arts. 66, 67):

α_1) Through two points A, B a straight line AB is determined, through which both points go. [...]
δ_1) Through two straight lines, which have a point in common, lay a plane.

α_2) Through two planes A, B a straight line AB is determined, in which the two planes intersect. [...]
δ_2) Find the point of intersection of two straight lines that lie in some plane.

2.4 Harmonic points and projective relationships. Among the 'n-corners' and 'n-edges' that von Staudt studied, quadrangles and quadrilaterals were the most interesting. Four coplanar points (vertices), no three of which were collinear, and their six connecting lines were said to form a complete quadrangle, if two straight lines which have no vertex in common, intersect in a diagonal point; there were three diagonal points which are not collinear. So each point of a quadrangular set was uniquely determined by the remaining points. Equivalent properties were true for the dual case of quadrilaterals. On this basis it was possible to define a 'harmonic configuration' ('*harmonisches Gebilde*'): four collinear points, of which two are a pair of opposite vertices of a complete quadrilateral and the other two are the intersections of their diagonals with the other two diagonals (art. 93).

Two fundamental one-dimensional configurations were defined as projective when the harmonic configuration of one of them could be set into one–one correspondence with the harmonic configuration of the other (art. 103). This definition of projectivity was very original since it did not involve the notion of distance; it could easily be extended to spatial systems (ch. 10). His main theorem stated that a projectivity was determined when three points on one of the straight lines and the corresponding three points on the other straight line were given (art. 110).

Further, von Staudt defined a collineation as a point–point relationship, transforming a straight line into a straight line while preserving harmonic configurations. A collineation was determined when two corresponding quadrilaterals were given (arts. 123–130). He also considered collineations in the case of spatial systems (arts. 132–136). The correlation is a second kind of a two-dimensional projectivity; there is a point-straight line relationship or the correlation between planes, using the relationship of four points in general position on a plane to four corresponding points or straight lines on another plane or on the same plane.

2.5 Involutions and polarity. The term 'involution' was due to Gerard Desargues in the 17th century. Von Staudt defined involution as a correspondence between the elements of one configuration to the other and vice versa, so that one returned to the original elements (art. 213). In this case the two configurations were called 'involutory'. There were involutory systems that were also collinear; non-collinear involutory systems were called 'polar systems' (art. 226). In a plane polar system every point in relation to a straight line was called a 'pole', and every straight line in relation to a point a 'polar'. Of special interest were the so-called 'polar triangles', which were determined by the polars of the vertices and the poles of the sides (arts. 236–243). He proved the following theorem: if there is a polar triangle in a plane and a point P, not located on any side of the triangle together with a straight line p, not coinciding with any of the vertices, then a polar system is determined (art. 237). This result laid the basis for a detailed theory of conic sections and second-order surfaces in general.

2.6 Conic sections. Von Staudt discussed conic sections either as loci of self-conjugate points or as an envelope of self-conjugate straight lines (arts. 247–248); thus a conic comprised not only the locus of a point but also the corresponding tangents. The conic was an ellipse when the curve did not have a common point with the vanishing line of the plane; it was a parabola, when the curve touched this vanishing line in one point; and it was a

hyperbola, when the curve intersected the vanishing line in two points. Every complete n-corner inscribed in a second-order curve defined a complete n-lateral, its sides being the polars of the vertices and the vertices being the poles of the sides (art. 250).

In the following pages von Staudt treated many details about the properties of triangles, 5-corners, and so on; for example, the vertices of those located on a second-order curve as well as special properties of second-order curves as projective relationships, common points, and tangents of two curves. In his last chapters he investigated polar systems in space and second-order surfaces, including a neat system of classifying types of singular points on space curves (points of inflection, cusps and horn angles) using plus and minus signs (art. 205).

3 ON VON STAUDT'S *BEITRÄGE ZUR GEOMETRIE DER LAGE*

About ten years later von Staudt presented a voluminous continuation under the above title, publishing it in three volumes in 1856, 1857 and 1860. He gave a projective foundation of the complex number field and developed the first complete theory of imaginary points, lines and planes in projective geometry [Fano, 1907, pt. 4]. In detail, he defined an elliptic point involution as a complex point; the same involution with the opposite sense was defined as the conjugate point. There were corresponding definitions of complex lines and planes. The 'cast' or 'throw' ('*Wurf*') was a number defined by four collinear points, four lines of a pencil, or four coaxial planes. He managed to create an algebra of throws, defining their sums and products, although the 'numbers' involved were just signs for representation that were defined in geometrical terms (0, 1 and ∞ are usually used for three points). He also devised a means of defining homogeneous coordinates of the points of space by means of his harmonic configuration [Torretti, 1978, 143–146].

Von Staudt's notions usually drew upon preceding ones in interesting and often novel ways. He was 'the most original and profound of the projective geometers of the German school. [. . .] his great passion [. . .] was for unity of method' [Coolidge, 1945, 61].

4 IMPACT

As we have stressed, von Staudt is not easy to read, and the reception of his books was not rapid. But their importance was recognized, especially for making clearer than had anyone else the gulf between metrical and projective notions. Arguably his best received contribution was the harmonic configuration (section 2.4); it became known as 'the quadrilateral construction', and was used in investigations of properties, especially invariance, of cross-ratios (a notion that von Staudt himself avoided, as we saw).

Von Staudt's work entered a rich melee of developments in geometries, where the non-Euclidean versions were also gaining much attention (§39.3–4, [Schönflies, 1909]; [Scriba and Schreiber, 2000, ch. 7]). An important representative of these joint concerns is Felix Klein (1849–1925), who removed von Staudt's dependence upon Euclid's parallel axiom, and added limit-points to his theory (compare [Klein, 1926, 132–140]; and §46.2). Von Staudt's work played a role in David Hilbert's first thoughts on geometry in the early 1890s (§55.3).

Among Italians Corrado Segre (1863–1924) extended projective geometry in various ways, including considering n-dimensional spaces and using bicomplex numbers i and j where $ij = ji$ and $i^2 = j^2 = -1$ [Segre, 1889]; while around 1897 Mario Pieri (1860–1913) axiomatised von Staudt's geometry, including axioms to handle continuity without invoking any new primitive notions [Marchisotto, 1995]. Another line of influence there was his use of collineation (section 2.4), which gave a possible new basis for projective geometry. Thereafter von Staudt became part of the heritage for Italian geometry ([Bottazzini, 2001]; compare §62).

However, supporters of synthetic geometry always had to confront algebraic projective geometers such as A.F. Möbius and then Julius Plücker, and their followers. From the 1820s they had been using algebra to express and study certain properties, such as construing non-intersecting curves as intersecting at points given by complex numbers, and indeed in due course reworking von Staudt's complex projective geometry itself [Kötter, 1898, esp. chs. 23, 33, 36 and 37; Fano, 1907].

The figure upon whom von Staudt's influence was most marked was Theodor Reye (1838–1919). He even gave his own book virtually the same title: *Die Geometrie der Lage* (2 volumes, 1866–1868). Reye's initial source was Karl Culmann (1821–1881), who guided him towards von Staudt's work, and he adopted parts of both men's theories. He treated linear manifolds of projective plane pencils and of collinear bundles or spaces, and founded point-series geometry. His book, much easier to read than von Staudt's, became so well known that it had five editions, up to 1923. In his foreword Reye emphasized the high quality, importance, and elegance of von Staudt's contributions to synthetic geometry.

BIBLIOGRAPHY

Böhmer, G. 1953. *Professor K. G. Chr. von Staudt. Ein Lebensbild*, Rothenburg ob der Tauber.

Bottazzini, U. 2001. 'I geometri italiani e il problema dei fondamenti (1889–1899)', *Bollettino dell'Unione Matematica Italiana, (8) 4A*, 281–329.

Coolidge, J.L. 1945. *A history of the conic sections and quadric surfaces*, Oxford: Oxford University Press.

Fano, G. 1907. 'Gegensatz von synthetischer und analytischer Geometrie', in *Encyklopädie der mathematischen Wissenschaften*, vol. 3, pt. 1, 221–288 (article IIIAB4a).

Freudenthal, H. 1974. 'The impact of von Staudt's foundations of geometry', in *For Dirk Struik*, Dordrecht: Reidel, 189–200. [Repr. in *Proceedings of the NATO Advanced Study Institute*, Bad Winsheim: 1981, 401–425.]

Hofmann, J.E. 1960. 'Karl Georg Christian von Staudt', *Veröffentlichungen der Gesellschaft für fränkische Geschichte, (7) 6*, 536–548. [Repr. in *Ausgewählte Schriften*, vol. 1, Hildesheim: Olms, 1990, 330–342.]

Klein, F. 1926. *Vorlesungen über die Entwicklung der Mathematik im 19. Jahrhundert*, vol. 1, Berlin: Springer. [Repr. New York: Chelsea, no date.]

Kötter, E. 1898. 'Die Entwickelung der synthetischen Geometrie von Monge bis auf Staudt (1847)', *Jahresbericht der Deutschen Mathematiker-Vereinigung, 5*, pt. 2, 486 pp.

Marchisotto, E.A. 1995. 'In the shadow of giants: the work of Mario Pieri in the foundations of mathematics', *History and philosophy of logic, 16*, 107–119.

Noether, M. 1923. 'Zur Erinnerung an Karl Georg Christian von Staudt', *Jahresbericht der Deutschen Mathematiker-Vereinigung, 32*, 97–119.

Pieri, M. 1904. 'Circa il teorema fondamentale di Staudt e i principi della geometria projettiva', *Atti della Reale Accademia delle Scienze di Torino*, 233–251.

Schönflies, A. 1909. 'Projektive Geometrie', in *Encyklopädie der mathematischen Wissenschaften*, vol. 3, pt. 1, 389–480 (article IIIAB5).

Scott, C.A. 1900. 'On von Staudt's Geometrie der Lage', *The mathematical gazette*, 307–314, 323–331, 363–370.

Scriba, C.J. and Schreiber, P. 2000. *5000 Jahre Geometrie. Geschichte Kulturen Menschen*, Berlin: Springer.

Segre, C. 1889. 'Carl Georg Christian von Staudt ed i suoi lavori', in *Geometria di posizione de Staudt*, Turin: Bocca, 1–17.

Steiner, J. 1832. *Systematische Entwicklung der Abhängigkeiten geometrischer Gestalten von einander*, Berlin: G. Fincke. [Repr. Leipzig: Engelsmann, 1896 (*Ostwalds Klassiker der exakten Wissenschaften*, nos. 85–86).]

Torretti, R. 1978. *Philosophy of geometry from Riemann to Poincaré*, Dordrecht: Reidel.

von Staudt, C. 1825. *Möglichst einfache Entwicklung des Gaußischen Theorems, die Theilung des Kreises betreffend*, Würzburg: Schulprogramm.

von Staudt, C. 1831. *Ueber die Kurven II. Ordnung*, Nürnberg: Schulprogramm.

CHAPTER 34

BERNHARD RIEMANN, THESIS ON THE THEORY OF FUNCTIONS OF A COMPLEX VARIABLE (1851)

Peter Ullrich

In his doctoral thesis, Riemann contributed to the foundations of complex function theory with the notion of a function of a complex variable and a discussion of the concept now called a Riemann surface. He characterized functions not by analytic expressions but by their properties, such as the nature and location of their singularities; and in application he gave his mapping theorem, that each simply connected (bounded) domain can conformally be mapped onto the unit circle. His approach met stiff competition from a quite different approach launched around the same time by Weierstrass.

First publication. *Grundlagen der allgemeine Theorie der Functionen einer veränderlichen complexen Grösse. Eine Abhandlung von B. Riemann*, Göttingen: Druck von Ernst August Huth, 1851. 32 pages. [Doctoral thesis at the Georg-August University at Göttingen.]

Manuscript. In Riemann's *Nachlass*, Göttingen University Library Archives.

Reprints. 1) Göttingen: Verlag von Adalbert Rente, 1867. 2) In *Gesammelte mathematische Werke* (ed. H. Weber and R. Dedekind), 1st ed., Leipzig: Teubner, 1876, 3–47. Also in 2nd ed., 1892, 3–48. Also in 3rd ed. (ed. R. Narasimhan), Berlin: Springer, 1990, 35–77.

Internet publication. www.emis.de/classics/-Riemann/index.html.

Italian translation. 'Fondamenti di una teorica generale delle funzione di una variable complesse' (trans. E. Betti), *Annali di matematica pura ed applicata, (1) 2* (1859), 288–304, 337–356.

French translation. In Riemann, *Oeuvres mathématiques* (ed. L. Laugel), Paris: Gauthier–Villars, 1898, 1–68.

English translation. 'Riemann's *Grundlagen*' (trans. J.C. Stillwell), Melbourne: Department of Mathematics, Monash University, 1978. [Omits the extensive table of contents.]

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

Related articles: Cauchy on complex-variable analysis (§28), Jacobi (§31), Poincaré (§48).

1 THE THEORY OF FUNCTIONS OF A COMPLEX VARIABLE BEFORE RIEMANN

1.1 Complex numbers. Complex numbers had entered mathematics in 16th-century Italian algebra in connection with the resolution of equations of the 3rd and 4th degree as purely formal expressions. Even in 1702 Gottfried Wilhelm Leibniz (1646–1716) would call them a 'subtle and wonderful resort of the divine spirit, a kind of hermaphrodite between existence and non-existence'. One motive to give these numbers a firmer place was the conviction that they could be used in order to find roots for any (non-constant) polynomial equation. This 'fundamental theorem of algebra', at first only hypothetical, was stated in 1629 by Albert Girard (1595–1632) and in 1637 by René Descartes (1596–1650) (§1).

The extension of non-algebraic functions to complex arguments, however, turned out to be a delicate enterprise. In the years 1712–1713 a controversy arose between Leibniz and Johann Bernoulli (1667–1748) on the values of the logarithms of negative and imaginary numbers, which was not resolved. Only in 1749 did Leonhard Euler (1707–1783) show that both Leibniz and Bernoulli were wrong with their (implicit) assumption that the rules for the calculation of logarithms of positive real numbers still hold for other arguments, and he showed that in the complex domain the logarithm becomes multi-valued. One year earlier, in his textbook *Introductio*, Euler had published his famous formula

$$e^{ix} = \cos x + i \sin x, \tag{1}$$

which sets up a connection between the exponential and the trigonometric functions for complex arguments (§13.2.4).

1.2 Elliptic and Abelian integrals. In 1742 Euler also gave an idea of a proof of the fundamental theorem of algebra. But what pushed forward even more the theory of functions of a complex variable was the research that he started in 1752 on elliptic integrals, namely, integrands of the form $\int 1/\sqrt{p(t)}\,dt$, where $p(t)$ is a polynomial of degree $\geqslant 3$. The primitives of these integrands could not be written in closed form, but it was known that, in the case of real coefficients in p, their inverse functions have a real period. The fact that they also have a second, complex period was not stated by Euler but found, independently, by Carl Friedrich Gauss (1777–1855) in 1797 and, later on, by Niels Henrik Abel (1802–1829).

Although Gauss did not publish his results, and Abel and Carl Gustav Jacobi (1804–1851) would at first reduce complex quantities to pairs of real ones, their later papers explicitly referred to complex functions of a complex variable that were expressed, for example, as infinite series or infinite products. In particular, these functions appeared when Abel and Jacobi generalized their studies to 'Abelian integrals' (as baptizised by Jacobi), that is, integrals of the type $\int q(f(t), t)\,dt$: here f is an algebraic function fulfilling a relation $P(f(t), t) \cong 0$, where P is a non-zero polynomial in two variables (compare

§31.5). Even more, in 1834 Jacobi felt forced to consider functions of several complex variables.

These Abelian integrals and the functions pertaining to them were a highly prestigious area of research during the 19th century. Both Karl Weierstrass and Bernhard Riemann set up their research on functions of complex variables in order to find means of handling such objects; for example, to obtain an analytic description of algebraic curves.

1.3 Augustin Louis Cauchy (1789–1857). Another source of the theory of complex functions was the wish to determine values of real integrals: (certain) pairs of real integrals were interpreted as one complex integral (for example, with the help of (1)).

Cauchy was the first to develop a theory of such complex integrals (§28). In his 'Mémoire sur les intégrales définies' (1814, published 1827) he considered only pairs of integrals with real limits; but in his booklet *Mémoire sur les intégrales définies prises entres des limites imaginaries* (1825) he defined complex path integrals in a way analogous to real ones and explained how this notion can be reduced to real path integrals. He also stated his theorem concerning the vanishing of the integral of a complex differentiable function along a closed path in this paper. In a paper of 1826 he extended this theory to the calculus of residues, which has always been the most important way to calculate real integrals with the help of complex path integrals.

By Cauchy's integral formula one can express the value of a complex (differentiable) function f at a point by means of an integral along a path running once around that point; for example, along the boundary $\partial B_r(c)$ of a disc $B_r(c)$ centered at the point c, namely:

$$f(c) = \frac{1}{2\pi i} \int_{\partial B_r(c)} \frac{f(z)}{z-c}\, dz. \tag{2}$$

This is first stated in a paper of 1831 'Sur la mécanique céleste et sur un nouveau calcul appelée calcul des limites', but became known to a wider audience only in 1841.

Cauchy did not study the multi-valuedness of the integrand (only that of the integral), and so missed a chance to use algebraic theory to understand the nature of branch points. But his younger compatriot Victor Puiseux (1820–1883) did enter this territory with papers of 1850 and 1851, just when Riemann was preparing his thesis, and produced a theory similar to analytic continuation [Brill and Noether, 1894, 190–202].

1.4 Carl Friedrich Gauss. Cauchy, however, was not the first with his researches. Already Euler, Pierre Simon Laplace (1749–1827) and Siméon Denis Poisson (1781–1840) studied integrals in the complex domain; but their efforts were rather heuristic. By contrast, a letter written to Friedrich Wilhelm Bessel (1784–1846) on 18 December 1811 shows that Gauss was then in possession of the Cauchy integral theorem (and also of the complex plane). During his lifetime, however, he did not publish his results either on elliptic functions or on integration in the complex domain.

Gauss had taken his doctorate in 1799 with a thesis on the fundamental theorem of algebra, in which he gave a proof, correct up to the current foundations of topology, the first one not attempting to construct the required root but just trying to show its existence.

During the next half century he published a further three proofs; the second one of 1816 is strict even by modern standards [Gauss, 1890]. Furthermore, in his paper [Gauss, 1825] on the making of maps, submitted for a competition at the Royal Academy of Sciences at Copenhagen, he proved that the maps of plane areas to plane areas that preserve angles and orientation locally around each point are precisely the complex differentiable functions of a complex argument; however, he did not use the language of complex analysis.

1.5 Karl Weierstrass (1815–1897). Although eleven years older than Riemann, Weierstrass started to publish his results on complex analysis in mathematical journals only three years after Riemann had written his doctoral thesis. They were both concerned with its applications to elliptic and Abelian integrals; indeed, Weierstrass had been inspired partly by the work on elliptic functions of his teacher at Münster University Christoph Gudermann [Manning, 1975]. His starting point for the definition of a function of a complex variable was power series (as they had been earlier in the works of Abel and Jacobi). As early as 1841–1842, but then unpublished, he was able to show first fundamental results for these series, for example his theorem on double series: that a uniformly convergent series of power series converges again to a power series. Furthermore, he explained even that early how one can define global functions from locally convergent power series by the process of analytic continuation [Ullrich, 2003].

At this stage of his life Weierstrass was a little-known school-teacher. But from the mid 1850s onwards he gained great fame as professor in Berlin University, with influential lecture courses and also a modest number of publications. We shall discuss the consequences in section 5, after we pick up Riemann's career.

2 BIOGRAPHY OF RIEMANN

In his short life Bernhard Riemann (1826–1866) contributed in fundamental ways to many areas of mathematics, especially real- and complex-variable analysis, analytic number theory and several areas of mathematical physics. Born to a Protestant minister in Breselenz south–east of Hamburg, he studied at Göttingen and Berlin Universities between 1846 and 1851. At Göttingen he came under the influence of Gauss: he became an extraordinary professor there in 1857 and full professor (actually in astronomy and mechanics) in 1859, when he succeeded J.P.G. Lejeune-Dirichlet (1805–1859), who himself had more or less succeeded Gauss. Always frail in health, Riemann spent periods from 1862 onwards in Italy with his family (he married that year and had a daughter), where he made important contacts with Italian mathematicians, as we shall see in section 6. Nevertheless, he died there in July 1866, a few months before his 40th birthday. There is no fitting biography, though the recollections [Dedekind, 1876] are precious. Many aspects of his work on analysis are reviewed in [Laugwitz, 1996]; see also [Bottazzini, 1986, ch. 6].

Riemann is represented in this book by his work prepared for doctorates at Göttingen University. For his *Habilitation*, the higher doctorate, in 1854 he wrote two essays, one on the foundations of geometry and the other on trigonometric series. He did not publish them, but they both made huge impacts when they appeared in 1867 under the editorship of his friend and fellow Gauss student Richard Dedekind (1831–1916) (§38, §39). Dedekind

seems also to have arranged the reprint at that time of the writing that is our concern here: the *Dissertation* that Riemann had submitted in 1851.

3 THE THESIS

Riemann seems to have chosen the topic of complex analysis for himself. At Berlin he was introduced by Dirichlet to the writings of the French analytic school, in particular Cauchy. There he also seems to have begun studying [Gauss, 1825]. The personal influence of Gauss at Göttingen, however, is not really clear: Dedekind [1876] gives the impression that Riemann had finished his thesis without any direct advice from Gauss, whereas Enrico Betti (1823–1892) claimed that Riemann formed the idea of his cuts in a private conversation with Gauss. At any rate, the only two sources cited in the thesis are [Gauss, 1825] and [Gauss, 1828]. Riemann appears to have defended it in December 1851.

3.1 Foundations of the complex-variable function. Compared to his forerunners Riemann offered a new foundation for the subject in his thesis, starting out not from an analytic expression but just from assuming that the complex function $w = u + vi$ of the complex variable $z = x + yi$ was differentiable (he said 'continuous', and ignored the question of existence of the derivative). The value of the derivative was given by

$$\frac{du + dvi}{dx + dyi} = \frac{(\frac{du}{dx} + \frac{dv}{dx}i)\,dx + (\frac{dv}{dy} - \frac{du}{dy}i)\,dy\,i}{dx + dy\,i}; \tag{3}$$

and it would be independent of the values of dx and dy if and only if the coefficients of dx and $dy\,i$ were equal, so that

$$\frac{du}{dx} = \frac{dv}{dy} \quad \text{and} \quad \frac{dv}{dx} = -\frac{du}{dy}. \tag{4}$$

(He wrote the differentials as 'd's; in the editions of his works they are rendered as '∂'s.) It also followed that both u and v satisfied Laplace's equation (arts. 1–4).

Riemann's approach brought complex function theory to well-known areas of real-variable mathematics: conformal mapping, especially in the paper [Gauss, 1825], which he cited; and all sorts of applications, especially potential theory. On the other hand, the equations in (4) implied that a complex (differentiable) function is already determined by its real part. They are now associated with him and Cauchy, but they play a still greater role in this new theory than they did in the Cauchy's, where their failure stimulated the study of singular integrals (§28, (7)). Interestingly, [Cauchy, 1851] had obtained (4) directly by letting z slide to the limit down the x and the y axes, shortly before approving for the *Académie des Sciences* a paper by Puiseux mentioned in section 1.3; Riemann did have some knowledge of recent work by Cauchy and Puiseux.

3.2 The Riemann surface. The next step was quite original (even if in some notes left behind by Gauss one can find first attempts in this direction). In order to consider elliptic

or Abelian integrals one has to handle algebraic and therefore often multi-valued functions. Instead of just treating each value separately in a purely analytical way, Riemann launched the idea of covering (parts of) the (finite) z plane A multiply by surfaces that are now named after him. A now took a finite surface ('*Fläche*') T over it, upon which complex points could sit. He imposed the condition that the (one-layer) parts of the surface ('*Flächentheile*') do not meet along a line. From this he deduced that on both sides of a line in the complex plane the number of parts of the surface lying above the plane is the same; he also discussed how the different parts fit together and form winding points ('*Windungspunkte*'), that is, points about which a neighbouring point would have to complete $m \geqslant 2$ revolutions on the surface before returning to its starting position. In order to lay the basis of an integration theory on these surfaces he introduced the concept of 'crosscuts' ('*Querschnitte*'); they divided the surface into simply connected ('*einfach zusammenhängende*') parts, upon which functions admitted unique integration (arts. 5–6). The connectivity ('*Ordnung des Zusammenhangs*') of a surface was the minimal number of crosscuts that would disconnect it.

The account is awe-inspiring but cryptic in the extreme; the topology was left entirely intuitive, and the scope of the approach uncertain, especially relative to the properties of functions that were then known. Exegeses of the theory would occupy mathematicians for generations, with an important stimulus coming in the 1890s from Henri Poincaré (compare §48) and in 1913 from Hermann Weyl (1885–1955) (section 5).

3.3 Functions and potentials. The rest of the thesis was devoted to the study of functions on these surfaces, their characterisation and also their construction (arts. 7–11). To this purpose, he relied on potential theory, especially relationships between surface and contour integrals, as handled by George Green (§30) and others from the 1830s and William Thomson a decade later (compare §40), and to some extent already in the electrodynamics produced in the 1820s by A.-M. Ampère, and in some results of Gauss. It is not clear how much of this earlier work Riemann knew, though he had definitely studied [Gauss, 1839]; in particular, Green's popularity dates largely from the 1850s. In any case Riemann conceived of these integrals in terms of (usually) continuous functions defined over surfaces and coverings and their boundaries.

Returning to (4), Riemann found conditions for the finitude and continuity of w over a surface; in particular, he proved the theorem on removable singularities named after him (art. 12), and also the consequences for behaviour when a discontinuity occurred at a value z' of z within a surface, when w could be expressed in terms of finite *inverse* power series in $(z - z')$ (arts. 12–14). From this he deduced the expansion of the surface at a branching point, as shown by Puiseux during the previous year. He was entering some recent Cauchyesque territory concerning singularities of functions, though Riemann made no mention of it; of course Cauchy used no notions corresponding to the surface. Among other results, he showed that if w satisfied (4) and was not a constant function over some surface, then it could not be constant along any line within it (art. 15).

Riemann then claimed that the integral

$$\int \left[\left(\frac{d\alpha}{dx} - \frac{d\beta}{dy} \right)^2 + \left(\frac{d\alpha}{dy} + \frac{d\beta}{dx} \right)^2 \right] dT \tag{5}$$

defined over any part of T, took a minimum value for functions α that were discontinuous only at isolated points at most, and were zero on the boundary of T (art. 16). Again he was in known territory; Green had made a similar claim in the 1820s, and later Riemann himself called it 'Dirichlet's principle' when he heard it used in Dirichlet's lectures. While he surpassed all predecessors in considering extensions of (5) to crosscuts and other variants (arts. 17–18), and the types of discontinuity to which complex functions may be suspect (art. 19), his use of (5) was to become a huge bone of contention in analysis, as we shall see in section 5.

Nearing the end, Riemann reflected upon his theory, especially that it did not rely on analytic expressions for functions of a complex variable but on their properties. This move let him reduce 'the number of determining components of a function', to quote the title given to the passage (art. 20) in the table of contents (which mostly stems from Riemann but is not printed in the first two publications of the thesis). As an application of his approach he gave a 'worked-out example', showing that two simply connected plane surfaces can always be made to correspond in such a way that each point of one corresponds continuously with its image in the other, and so that corresponding parts are 'similar in the small', or conformal; tacitly assuming that the complex plane was not under consideration, he took the unit circle around the origin as the reference surface for what is nowadays called the 'Riemann mapping theorem' (art. 21).

4 RIEMANN'S PUBLICATIONS FROM 1857

The thesis had at least one contented reader: university examiner Gauss. who wrote a terse but positive report [Remmert, 1993; Laugwitz, 1996, 124]. But there may not have been many more readers: although a *Dissertation* was a printed booklet, it was not usually published or publicised in the normal way; the candidate had to pay for the print-run, and sales and marketing were executed on an infinitesimal scale. So the first printing of Riemann's thesis consisted only of the obligatory copies he had to hand in at Göttingen University, and a few copies for personal use.

But when Riemann passed his *Habilitation* examination in 1854, he gained the right to lecture at the university, and during the remaining 12 years of his life his teaching included courses on complex analysis, function theory in general, elliptic functions, and differential equations. The contents of the thesis became known to a wider public especially from 1857. One paper published that year dealt with 'Gauss's series', that is, the hypergeometric function, cast in complex variables [*Works*, 67–87], which Gauss had handled in 1813; Riemann extended his treatment to an axiomatically defined class of functions that he called '*P*'. Also that year he published three short notes in *Journal für die reine und angewandte Mathematik* on the main ideas of his thesis: the concept of complex function and (Riemann) surface; integration on multiply connected domains; and the determination and especially the definition of functions by given conditions [*Works*, 88–100]. He followed straight on with a long paper on Abelian functions that made wide use of the thesis [*Works*, 100–144].

Two years later Riemann plunged into analytic number theory, and surfaced with his most famous conjecture, still unresolved. It concerns the location in the complex plane of the zeroes of the zeta function [*Works*, 145–153].

5 THE REACTION OF WEIERSTRASS

In his main article of 1857 Riemann indicated how to solve the problems concerning Abelian integrals; he knew, from an article of 1854, that Weierstrass was also working on them. Weierstrass had prepared another paper on this topic, but he withdrew it from the printer, not only because Riemann was prior to him but also since it was by no means obvious how to translate the results found by one of them into the language of the other. According to a story told by Arnold Sommerfeld (1868–1951), Weierstrass started an intense study of Riemann's doctoral thesis only in the 1870s, and he found difficulties with the latter's physical intuition [Bottazzini, 2002; Ullrich, 2003].

The difference between the two approaches is particularly evident in their areas of common concern. For example, Weierstrass based his notion of analytic function on the principle of analytic continuation; Riemann, on the other hand, did use it sometimes, but only as a technical tool [Neuenschwander, 1980]. To handle multi-valuedness, especially regarding Abelian functions, Weierstrass reacted against Riemann's surfaces and created a theory of 'analytic configurations', where he took collections of pairs of variables related by such functions and examined their various properties, sometimes using parametric representations of the variables [Ullrich, 2003].

In addition to such dissimilarities, there was one direct refutation, concerning Dirichlet's principle. Rumours against its use had been around at least as early as the late 1850s, but Riemann did not attach great importance to its specific use in his existence proofs. However, in 1870 Weierstrass made public a counter-example to the assumption that integrals like (5) took a minimum among the possible functions, so that arguments based upon it were not secure. The consequences were felt most severely among practitioners of the principle in potential theory in mathematical physics, but it also bore upon complex-variable analysis; indeed, a further bringer of bad news was Riemann's student Emil Prym.

6 THE POSITIVE RECEPTION OF RIEMANN'S THESIS

From the late 1850s Riemann began to gain followers abroad, in particular in Italy where the state of his health forced him to retreat [Bottazzini, 1977]. His closest context there, indeed friend, was Betti, who published a translation of the thesis in 1859 and began to explore the topological wonders of the surfaces. Another fan was Felice Casorati (1835–1890), whose *Teorica delle funzioni di variabili complesse* of 1868 included not only modern theories but also a substantial historical account.

The most influential supporter of Riemann's approach was the Göttingen professor Felix Klein (1849–1925), an enthusiast for geometry—for example, the newly fashionable non-Euclidean kinds (compare §42). His own researches in function theory drew on several aspects of Riemann's work or their consequences as drawn by others [Gray, 1986]. In addition to his technical work he published a short book explaining to a broader mathematical audience Riemann's treatment of algebraic functions [Klein, 1882].

More general publicity was given during the 1890s by the newly formed *Deutsche Mathematiker-Vereinigung*, which began the admirable practice of including survey articles in its yearly *Jahresbericht*. The third volume contained a book-length account of

algebraic functions, and Riemann featured more than any other single figure (including Weierstrass), with detailed summaries of the thesis and the 1857 papers [Brill and Noether, 1894, pts. 3–4]. They also reported on some Riemann manuscripts [ch. 5C]; and in 1902 one of the authors, Max Noether, co-edited a supplement to Riemann's collected works, which Dedekind and Heinrich Weber had produced in editions of 1876 and 1892, where a lot of pertinent manuscript material was published, especially some lecture courses.

Meanwhile, Klein's influence is evident again when from the mid 1890s he began to organise the *Encyklopädie der mathematischen Wissenschaften*. The second Part of the second 'volume' was devoted to complex analysis, and its opening article was given to the American mathematician W.F. Osgood (1862–1943), who had written his *Dissertation* at Göttingen in 1890 on Abelian functions. Osgood [1901] exemplifies well the pragmatic view of the varying approaches that many complex analysts seem to have adopted; using each one according to the context as fitted mood or method. He started with Cauchy on series expansions but soon brought in Riemann surfaces, and provided a substantial part on 'the geometric function theory'; but he then followed with a comparable one on Weierstrass's approach. The final part dealt with functions of several complex variables, where Weierstrass and his followers were the only ones that had the adequate tools at hand at that time.

In his article Osgood announced that he was preparing a monograph on the subject; a large *Lehrbuch der Funktionentheorie* appeared in 1907, and thereafter in four further editions up to 1928, the fourth onwards appearing in two volumes. Once again Riemann fared well, with a substantial chapter on his surface, and other topics featured well. One was the 'logarithmic potential', which was the name that had been given since the 1870s to solutions of Laplace's equation in the plane.

Riemann's use of the Dirichlet principle was seen in a brighter light at that time. Klein had already tried to corroborate Riemann's results by appealing to physical intuition. Now Henri Poincaré (1854–1912) proved the solvability of Dirichlet's problem under fairly general conditions [Poincaré, 1899], and David Hilbert (1862–1943) finally proved a precise version of Dirichlet's principle which is sufficiently general to allow for the usual function-theoretic applications [Hilbert, 1900, 1904].

The final vindication of Riemann's thesis, however, was given by Weyl in a lecture course at Göttingen University from which he drew book *Die Idee der Riemannsche Fläche* [Weyl, 1913]. Using the recently developed set-theoretic topology (compare §46), he gave a formal definition of a Riemann surface, even a re-interpretation or even a re-writing of Riemann's original ideas; the essays in the recent edition of [Weyl, 1913] depict the large distance between Riemann's thesis and its modern reading.

7 A COMPLEX OF THEORIES

From around 1880 or so three traditions in complex analysis, especially function theory, were evident [Markushevich, 1955, 1996; Neuenschwander, 1981]. Cauchy's theory of the integral was still powerful, as a growing body of theorems about contours and residues inside or on them; but his general view of the functions was becoming subsumed under either

Weierstrass's strict use and control of power series, or Riemann's 'geometric imagination' (K. Weierstrass) about surfaces and their cuts. Even the French mathematicians began to take serious note of these German alternatives [Neuenschwander, 1998].

Seen from the mathematical point of view, these approaches are equivalent, supposing that one has a correct proof of the Cauchy integral formula (2) for differentiable complex functions. This was not the case until in 1883 Edouard Goursat (1858–1936) gave a sound proof. After some simplifications by Goursat himself in 1899 and by Alfred Pringsheim (1850–1941) in 1901, it was reduced to the fact that the path integral along the boundary of a triangle vanishes for each function which is complex differentiable throughout the closure of the triangle. Therefore from that time onwards a fusion of all approaches to complex analytic functions—Cauchy's, Weierstrass's or Riemann's—was possible as far as mathematical arguments were concerned. It took, however, still some decades in the 20th century until this fusion was completed also in textbooks. Even the posthumous one of Adolf Hurwitz (1859–1919) edited by Richard Courant (1888–1972) shows the difference; Hurwitz had followed Weierstrass, while Courant was with Riemann [Hurwitz, 1922].

8 CONCLUDING REMARK

The issues between Riemann's and Weierstrass's approaches are quite stark; few other branches of mathematics show such wide divergences in possible manners of basic treatment. Klein captured the dissimilarity beautifully [1926, 246]:

> Riemann is the man with the shining intuition. Through his all-embracing genius he surpasses all his contemporaries. When his interest is awakened, he starts afresh, without being led astray by intuition and without acknowledging the coercive pressure of systematisation.
> Weierstrass is in the first place a logician; he advances slowly, systematically, step by step. Where he works, he strikes for the definitive form.

BIBLIOGRAPHY

Bottazzini, U. 1977. 'Riemann's Einfluss auf E. Betti und F. Casorati', *Archive for history of exact sciences, 18*, 27–37.

Bottazzini, U. 1986. *The higher calculus: real and complex analysis from Euler to Weierstrass*, New York: Springer.

Bottazzini, U. 2002. ' "Algebraic truths" *vs* "geometric fantasies": Weierstrass' response to Riemann', in *Proceedings of the International Congress of Mathematicians, ICM 2002*, vol. 3, Beijing: Higher Education Press, 923–934.

Brill, A. von and Noether, M. 1894. 'Die Entwicklung der Theorie der algebraischen Functionen', *Jahresbericht der Deutschen Mathematiker-Vereinigung, 3*, i–xxiii, 109–566.

Cauchy, A.L. 1851. 'Sur les fonctions de variables imaginaires', *Comptes rendus de l'Académie des Sciences, 32*, 160–162. [Repr. in *Oeuvres complètes*, ser. 1, vol. 11, 301–304. See also pp. 276–284/325–335 on Puiseux.]

Cooke, R. 1989. 'Abel's theorem', in D. Rowe and J. McCleary (eds.), *History of modern mathematics*, vol. 2, Boston: Academic Press, 389–421.

Hurwitz, A. 1922. *Vorlesungen über allgemeine Funktionentheorie und elliptische Funktionen*, Berlin: Springer.

Dedekind, R. 1876. 'Riemanns Lebenslauf', in Riemann *Works*, 1st ed., 507–526. [Repr. in 2nd ed. (1892), 539–558.]

Gauss, C.F. 1825. 'Analytische Auflösung der Aufgabe [...]', *Astronomische Abhandlung, 3*, 1–30. [Repr. in *Werke*, vol. 4, 189–216.]

Gauss, C.F. 1828. 'Disquisitiones generales circa superficies curvas', *Commentationes Societatis Regiae Scientiarum Gottingensis recentiores, 6* (1823–1827), classis mathematicae, 99–146. [Repr. in *Werke*, vol. 4, 217–258.]

Gauss, C.F. 1839. 'Allgemeine Lehrsätze in Beziehung auf die im verkehrten Verhältnisse des Quadrats der Entfernung wirkenden Anziehungs- und Abstossungs-Kräfte', in *Resultate aus den Beobachtungen des magnetischen Vereins*, 1–51. [Repr. in *Werke*, vol. 5, 195–242.]

Gauss, C.F. 1890. *Die vier Gauss'schen Beweise für die Zerlegung ganzer algebraischer Functionen in reelle Factoren ersten und zweiten Grades* (*1799–1849*) (ed. E. Netto), Leipzig: Engelmann (*Ostwald's Klassiker der exakten Wissenschaften*, no. 14).

Gray, 1986. *Linear differential equations and group theory from Riemann to Poincaré*, Basel: Birkhäuser.

Hilbert, D. 1900. 'Über das Dirichletsche Prinzip', *Jahresbericht der Deutschen Mathematiker-Vereinigung, 8*, 184–188. Also in *Journal für die reine und angewandte Mathematik, 129* (1905), 63–67. [Repr. in *Abhandlungen*, vol. 3, 10–14.]

Hilbert, D. 1904. 'Über das Dirichletsche Prinzip', *Mathematische Annalen, 59*, 161–186. [Repr. in *Abhandlungen*, vol. 3, 15–37.]

Klein, F. 1882. *Über Riemann's Theorie der algebraischen Functionen und ihrer Integrale*, Leipzig: Teubner. [English trans.: *On Riemann's theory of algebraic functions and their integrals*, Cambridge: Macmillan and Bowes, 1893.]

Klein, F. 1926. *Vorlesungen über die Entwicklung der Mathematik im 19. Jahrhundert*, vol. 1, Berlin: Springer. [Repr. New York: Chelsea, no date.]

Laugwitz, D. 1996. *Bernhard Riemann*, Basel: Birkhäuser. [In German. English trans. 1999.]

Manning, K.R. 1975. 'The emergence of the Weierstrassian approach to complex analysis', *Archive for history of exact sciences, 14*, 297–383.

Markuschewitsch, A.I. 1955. *Skizze zur Geschichte der analytischen Funktionen*, Berlin: Deutscher Verlag der Wissenschaften.

Markushevich, A.I. 1996. 'Analytic functions', in A.N. Kolmogorov and A.P. Yushkevich (eds.), *Mathematics of the 19th century. Geometry. Analytic function theory*, Basel: Birkhäuser, 119–272.

Neuenschwander, E. 1980. 'Riemann und das „Weierstrasssche" Prinzip der analytischen Fortsetzung durch Potenzreihen', *Jahresbericht der Deutschen Mathematiker-Vereinigung, 82*, 1–11.

Neuenschwander, E. 1981. 'Studies in the history of complex function theory II: interactions between the French school, Riemann, and Weierstrass', *Bulletin of the American Mathematical Society, 5*, 87–105.

Neuenschwander, E. 1998. 'Documenting Riemann's impact on the theory of complex functions', *The mathematical intelligencer, 20*, no. 3, 19–26.

Osgood, W.F. 1901. 'Allgemeine Theorie der analytischen Funkionen a) einer und b) mehrerer complexen Grössen', in *Encyklopädie der mathematischen Wissenschaften*, vol. 2, pt. 2, 1–114 (article IIB1). [See also the later articles in this part for several specific topics.]

Poincaré, H. 1899. *Théorie du potentiel Newtonien*, Paris: Gauthier–Villars.

Remmert, R. 1993. 'The Riemann-file Nr. 135 of the Philosophische Fakultät of the Georgia Augusta at Göttingen', *The mathematical intelligencer, 15*, no. 3, 44–48.

Riemann, G.F.B. *Works. Gesammelte mathematische Werke* (ed. H. Weber and R. Dedekind) 1st ed., Leipzig: Teubner, 1876. [2nd ed. 1892 (cited here); repr. with the 1902 supplement New York: Dover, 1953. Repr. as 3rd ed. with new bibliographical material, Berlin: Springer, 1990.]

Ullrich, P. 2003. 'Die Weierstrassschen „analytischen Gebilde": Alternativen zu Riemanns „Flächen" und Vorboten der komplexen Räume', *Jahresbericht der Deutschen Mathematiker-Vereinigung, 105*, 30–59.

Weyl, H. 1913. *Die Idee der Riemannschen Fläche*, Leipzig: Teubner. [2nd ed. 1923 (repr. New York: Chelsea 1951). 3rd ed. Stuttgart: Teubner, 1955. New ed. by R. Remmert, Leipzig and Stuttgart: Teubner, 1997.]

CHAPTER 35

WILLIAM ROWAN HAMILTON, *LECTURES ON QUATERNIONS* (1853)

Albert C. Lewis

This, the first book devoted to quaternions, appeared ten years after their discovery by Hamilton. Later, many of his most useful concepts were separated from their quaternion context and were reformulated as a part of vector analysis. The key work in this transformation was E.B. Wilson's *Vector analysis* (1901).

First publication. *Lectures on quaternions containing a systematic statement of a new mathematical method of which the principles were communicated in 1843 to the Royal Irish Academy*: *and which has since formed the subject of successive courses of lectures delivered in 1848 and subsequent years, in the halls of Trinity College, Dublin*: *with numerous illustrative diagrams, and with some geometrical and physical applications*, Dublin: Hodges and Smith; London: Whittaker and Co.; Cambridge: Macmillan and Co., 1853. lxxii + 736 pages.

Reprint. Cornell Library Digital Collections (http://historical.library.cornell.edu). [Preface only in *Mathematical papers*, vol. 3.]

Related articles: Grassmann (§32), Heaviside (§49).

1 FROM PRODIGY TO SAGE

Sir William Rowan Hamilton (1805–1865), born and raised in Ireland, was one of the most brilliant students to have passed through Trinity College, Dublin. He mastered many languages, ancient and modern. He appears to have been largely self-taught in mathematics though guided by a tutor. At Trinity from 1823 to 1827 he was exposed to the newest mathematics that emanated mainly from France, especially by P.S. Laplace, J.L. Lagrange, S.D. Poisson, and S.F. Lacroix. He had barely completed his intended program of studies when he was offered the prestigious appointment of Astronomer Royal of Ireland. Though he was not inclined to the practical aspects required for the job, he gained the aid of his four

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

sisters who lived with him at the Dunsink Observatory and who managed the household and performed many if not most of the astronomical duties; and the post provided more time for his mathematical researches than the conventional teaching position for which he originally aimed.

Hamilton's initial fame beyond Dublin came from work on systems of light rays done already in 1824. The experimental verification of his prediction of the conical refraction of light in 1833 brought him the highest honors, including a knighthood in 1835. Another enduring work from this early period of his life was what came to be called the 'Hamiltonian function', which has proven of fundamental importance in physics. It was also during this period that he began the pursuit of the elusive triple number system that led to his discovery of quaternions in 1843 and to the book featured here. Though quaternions, with their promise of so many fruitful applications, was to take much of his attention in later life, he devoted himself to many other matters.

From 1837 to 1846 Hamilton was president of the Royal Irish Academy. Far from being an exclusively honorary post, this entailed an intimate involvement in its administration and in the development of wide-ranging projects relating to Irish history and culture. After he resigned this post he was widely praised for his achievements in the Academy, not least for his diplomatic skill at resolving disputes among members. He was spurred by the discovery of the planet Neptune in 1846 to study perturbation theory. One result was his invention of the hodograph, an elegant geometrical representation of planetary paths which was taken up with some interest by William Thomson (Lord Kelvin) among others. Hamilton later learned that A.F. Möbius had discovered the notion earlier, as he acknowledged in the *Lectures* (p. 614).

In 1856 Hamilton became interested in an entirely different subject, which he termed 'the Icosian Calculus'. This was an algebra capable of describing the paths connecting the vertices of a dodecahedron. From this came the general idea of determining what have come to be known as 'Hamilton circuits'. Further details can be found in the principal secondary sources on his life [Graves, 1882–1891] and [Hankins, 1980].

2 THE ORIGIN OF QUATERNIONS

During the 1830s Hamilton maintained an interest in a problem that a number of mathematicians regarded as one of the most important unsolved issues of the time: How can the system of number pairs, represented by complex numbers, be extended to triples of numbers in such a way as to preserve the same operational properties? For example, a complex number $z = a + bi$ can be represented in the Euclidean plane by the directed line segment from the origin to the point with real number coordinates (a, b). On multiplying z by i $(= \sqrt{-1})$, the result would be the segment from the origin to $(-b, a)$ which can be regarded as the result of rotating the original segment 90° counterclockwise about the origin. The problem could thus be put in a geometrical and somewhat more general fashion: How can this mathematical operation, represented by rotation about a point in the plane, be extended to rotation about a line in three dimensions? Expressed this way the answer turns out to be that four numbers, not three, are required.

In his early work Hamilton assumed, not unnaturally, that the task was one of finding the appropriate system of triples of numbers, and it was only after many unsuccessful ef-

forts that the possibility of a quadruple of numbers presented itself. In one of the famous moments in the history of mathematics, the idea came to him—as he carefully recorded immediately afterwards—while on Brougham Bridge in Dublin, on his way to attend a Council meeting of the Academy on 16 October 1843. His construction consisted of adjoining three entities, i, j, k, to the real numbers. These can be multiplied according to the rules which he jotted down at the time in a notebook:

$$i^2 = j^2 = k^2 = -1, \quad ij = k, \quad jk = i, \quad ki = j \quad \text{and} \quad ijk = -1. \tag{1}$$

The hypercomplex number can be formed as $a + bi + cj + dk$, where a, b, c, and d are real numbers. Though this new number, Hamilton's quaternion, suggests that it might best be suited for a four-dimensional spatial representation—and indeed in the early 20th century there were attempts to make use of it in relativity theory—Hamilton himself exploited it for a wide range of three-dimensional applications, his original motivation.

Before taking up Hamilton's development of quaternions in the Lectures proper, mention should be made of his path of discovery which he described at length and in several places, including in the Preface of the *Lectures.* In his view the discovery is intimately tied to his notion of algebra as the science of pure time. As he describes it in the Preface, he tended to approach the whole subject less in a 'symbolical' fashion than in a 'scientific' fashion. Influenced by 'the Kantian parallelism between the *intuitions* of Time and Space', and by geometry as the science of space, he felt that viewing algebra as the science of pure time had a high suggestive value that could easily lead to a purely symbolical calculus if and when one chose to follow that symbolical route. Hamilton's detailed documentation of his creative path has been the basis of several historical analyses, some of which attempt to use it to draw lessons about the nature of mathematical discovery in general [Hankins, 1980, ch. 6; Pickering, 1995, ch. 4].

There are a number of predecessors for Hamilton's work whom he acknowledges in the Preface. The most significant one for later developments is 'the very original and remarkable work' of H.G. Grassmann whose *Ausdehnungslehre* or calculus of extension of 1844 (§32) he read just as the *Lectures* was being completed. Hamilton noted that, though Grassmann had a non-commutative multiplication of directed lines, he was not in possession of quaternions since he admitted to not succeeding in extending the complex numbers to three dimensions or in building a theory of angles in space (Preface, p. 62). It seems that Hamilton had quaternions predominately on his mind as he read Grassmann and overlooked the fact that, as later readers recognized, Grassmann's response to these issues was considerably more general than his own.

3 THE *LECTURES*

After the inspiration of October 1843 Hamilton published a number of very substantial papers over the next ten years describing the new entities, including two series of papers, one in eighteen installments in the *Philosophical magazine* and another, left incomplete after ten installments, in the *Cambridge and Dublin mathematical journal.* (These papers are reprinted in the edition [Hamilton, *Papers*].) In 1848 he conducted a series of lectures at Trinity College, Dublin, and these formed the basis for his *Lectures* volume of 1853. While

this was a most productive time for Hamilton with respect to developing and propagating the quaternions, he was also undergoing some traumatic personal experiences: his favorite sister, Eliza, died in 1851, and he had been following closely the health and well being of his greatest love in life, Catherine Disney Barlow, who attempted suicide in 1848. One of his biographers, Thomas Hankins, paints a picture of someone who was clearly deeply affected by these and other personal setbacks but who could appear totally unfazed by them when it came to carrying on with his work.

The contents of Hamilton's *Lectures* are summarised in Table 1. His path from consideration of progression in time, through working with number triplets, to quaternions, is given with substantial technical detail in the Preface. In fact, the Lectures themselves have been, as he puts it, 'drawn up in a more popular style than this Preface' and were intended, at least initially, to be fairly faithful to what was actually presented by Hamilton 'in successive years, in the Halls of this University'. However, as the table of contents reveals, the lengths of the 'Lectures' increased steadily, culminating in Lecture VII, which is over 300 pages long. As he admitted, substantially more 'calculation' was added than would actually have been presented in the lecture hall. Nevertheless, he maintains that 'something of

Table 1. Contents by Lectures of Hamilton's book. Parentheses around the Preface's page numbers distinguish them from the main body. Square brackets indicate an unnumbered page.

Lecture	**Page**	**Sect.**	**Art.**	**Contents (sample topics)**
Preface, pp. ([1])–(64)				Time, number triplets, quaternions.
Contents, pp. [ix]–lxxi				[No pages correspond to numbers i to viii.]
Lec. I	1	i	1	Addition and subtraction of lines and points.
Lec. II	33	vi	37	Multiplication and division in geometry; squares and products of i, j, k.
Lec. III	74	xi	79	The quaternion; tensor and versor.
Lec. IV	130	xxvi	121	Powers and roots of quaternions; $\sqrt{-1}$ as a partially indeterminate symbol.
Lec. V	186	xxxvi	175	Multiplication of three lines in space; value of ijk and kji.
Lec. VI	241	xlvi	251	General associative property of multiplication; spherical representations.
Lec. VII	381	lxi	394	Addition and subtraction; distributive principle of multiplication. [End cxvii, 689.]
App. A	701			Gauche (i.e. non-planar) polygons inscribed in second-order surfaces. [Paper published in 1850.]
App. B	717			Gauche polygons inscribed in second-order surfaces. [Paper published in 1849.]
App. C	731			A 'rapid outline of the quaternion analysis'. [End 736.]
Errata, unnumbered leaf				Thirty-seven errata, most quite minor.

the style of actual lecturing has been here and there retained' throughout (Preface, p. 63). Apparently to help the reader, if not the author, to gain control over the growing mass of material, Section numbers came into play at a later stage, 'too late to be incorporated into the text' as Hamilton writes (p. lxxi). Thus the 117 Sections, which act as an intermediate level of division between the 7 Lectures and 689 articles, occur only in the Table. There is no index, but the copious table of contents serves as an analytic guide. This glimpse into what was evidently a struggle to keep the book under control illustrates a general trait: as biographer Hankins put it, 'Hamilton could not keep his published works on quaternions within reasonable bounds' [Hankins, 1980, 365].

Hamilton's previous course of lectures was an introduction to astronomy and his first lecture in this new series—evidently the only one in the *Lectures* to adhere at all closely to what he might have actually said in the lecture hall—makes the transition by drawing upon relative positions of planets to introduce the notion of a difference of points as an ordinal expression of relative position. 'And because, according to the foregoing illustrations, this sign or mark (Minus) directs us to DRAW, or to conceive as drawn, *a straight line connecting the two points*, which are proposed to be compared as to their relative positions, it might, perhaps, on this account be called the SIGN OF TRACTION' (p. 10). This sentence succinctly exhibits something of Hamilton's style. Also, it shows his care in giving new concepts correspondingly new names even at the risk of overloading the reader. In this case, in the interest of reducing the number of new terms, 'subtraction' is soon used instead. In the next dozen pages, however, more terms come into play—almost at the rate of one per page—such as 'vection', 'revection', 'provection', and 'transvection' to describe various possible motions of a point along a line. Besides the subtraction of two points, the other key notion of the first Lecture is the geometrical meaning of sum of a line and a point. If $B - A$ is conceived as the line from A to B, then the sum $(B - A) + A$ results in the point B.

The second Lecture concerns a general division and multiplication that are the analytic and synthetic cardinal operations corresponding to the ordinal operations of subtraction and addition introduced in the previous Lecture. The term 'cardinal' comes from the analogy that, given an expression such as $\beta = n + n$ (Greek letters will represent directed line segments), we can ordinarily regard the quotient $\beta \div n$ as the cardinal number 2. The defining expressions are: $\beta \div \alpha = q$ and $q \times \alpha = \beta$. This last equation shows how the quotient q can be regarded as an operator that produces one directed line segment from another. If q is a 'tensor', or signless number, then it affects only the length of α. If q is a sign ($+$ or $-$) it changes the direction of α. If it is a real number then q may have the effect of changing both the direction and length. If it is a 'vector-unit' (or 'quadrantal versor'), i, j, k, then the effect is to turn α right-handedly through 90 degrees in a plane perpendicular to the vector-unit. Hamilton points out that a multiplication of a vector-unit, say i, by itself results in a rotation of 180°, i.e. the same as multiplying by -1 (reversing its direction) or $i^2 = -1$.

Continuing with further examples in Lecture III, Hamilton broadens the conception of multiplication to include any two vectors (directed line segments) and also introduces exponentiation of vectors. He shows that these operations, as well as the quotient, q, of two vectors described above, can be characterized by four numbers, namely the tensor (a pure number or scalar, written as Tq) and three directions. This entitles the result to be called a

quaternion. Most of the previous examples can thus be recognized at this stage as special cases of quaternions.

Lectures IV and V begin to reveal the nature of the quaternion itself and further properties of quaternion multiplication. Demonstrations of the non-commutative and associative properties of multiplication are given for more general cases, though not yet in the most general case. (A demonstration of its distributivity over addition is also promised, but this is delayed until Lecture VII as Hamilton feels the need to explore multiplication more before taking up addition.) His arguments throughout are geometrical. He makes use of 'arcual constructions' in which 'representative arcs' and 'representative angles' on the sphere are 'intimately connected' with versors though 'distinct from them'. This approach may have been deemed particularly appropriate since these Lectures followed on ones devoted to astronomy where spherical geometry figured prominently. Thus the product of two versors is represented 'by the external vertical angle of a spherical triangle, whose base angles, taken in a determined order, represent those two versors themselves' (p. 385). In Lecture VI, for example, he describes the 'symbol of operation' $q(\ )q^{-1}$ 'in which q may be said to be the *operating quaternion*, as denoting the operation of causing the arc which represents the *operand quaternion*, and whose symbol is supposed to be inserted within the parentheses, to *move along the* DOUBLED ARC of the operator, without any change of either *length* or *inclination* (like the equator on the ecliptic in precession)' (p. xxviii). Lecture VI also contains the general proof of associativity of multiplication.

Finally Lecture VII introduces addition of the various entities thus far introduced. For example, the addition of a scalar and a vector is shown to be a quaternion. Hamilton first justifies this for the case of a unit scalar and a unit vector by considering $1 + k$. If each term is multiplied on the right by i the results are i and $ki = j$. Thus

$$1 + k = (i + ki) \div i = (i + j) \div i, \tag{2}$$

and this last expression has a meaning that has been already established, namely the quotient of two vectors which has been shown to be a quaternion (pp. 387–388). Conversely, it is shown that a quaternion, q, is decomposable into a scalar and a vector. The operations of taking the scalar and vector are written as Sq and Vq respectively. The vector of the product of two vectors is shown to have length equal to the area of the parallelogram formed by the two vectors and a direction perpendicular to the plane of the parallelogram (the modern cross product). The product changes sign if the factors are interchanged (pp. 416–417). It is only in art. 450 that the 'quadronomial form' is formally introduced whereby a quaternion can be expressed in general as a sum of four terms, $q = w + ix + jy + kz$, where w, x, y, and z are numbers. If a second quaternion is written as $q' = w' + ix' + jy' + kz'$ then their sum or difference is formed by the following:

$$q \pm q' = (w \pm w') + i(x \pm x') + j(y \pm y') + k(z \pm z'). \tag{3}$$

The two quaternions are equal if and only if the system of four equations holds: $w = w'$, $x = x'$, $y = y'$, and $z = z'$. In spite of this introduction of an algebraic approach, the presentation remains geometrically oriented. Many illustrations of the use of quaternions to represent geometric figures and their intersections, in particular the conics and their surfaces

of revolution such as the ellipsoid, are given. Relations to trigonometry and goniometry (functions of angles) are developed.

In the middle of Lecture VII is a brief account of the form in which Hamilton originally discovered and expressed the quaternions (arts. 530–536). The 'quadronomial' form was central in that early stage and the geometrical connections developed gradually from it. He cites here the contribution made by his friend J.T. Graves in propagating some of these early results. Functions of quaternions are also introduced in Lecture VII, including exponential and logarithmic. Hamilton describes what is now referred to as the nabla or del operator, which he had introduced in 1846, and the Laplace operator:

$$\nabla = i\frac{\mathrm{d}}{\mathrm{d}x} + j\frac{\mathrm{d}}{\mathrm{d}y} + k\frac{\mathrm{d}}{\mathrm{d}z}, \quad \text{and} \quad \nabla^2 = -\left(\frac{\mathrm{d}^2}{\mathrm{d}x^2} + \frac{\mathrm{d}^2}{\mathrm{d}y^2} + \frac{\mathrm{d}^2}{\mathrm{d}z^2}\right). \tag{4}$$

Hamilton constructs what he called 'biquaternions', entities of the form $q' + \sqrt{-1}q''$ where q' and q'' are '*real quaternions*' and the $\sqrt{-1}$ is 'the *old* and *ordinary imaginary of algebra*' (p. 638). Two non-zero biquaternions may have a product of zero. (The term 'biquaternion' was to be used later by W.K. Clifford in a different sense.) The Lecture includes discussion of connections of quaternions with coordinates, determinants, trigonometry, series, linear and quadratic equations, differentials, integration, and continued fractions. Additional examples are given of quaternion representations of the differential geometry of curves and surfaces in three-dimensions.

4 RECEPTION AND SUBSEQUENT DEVELOPMENT

The *Lectures* probably did not sell many copies, but at least Hamilton had his printing costs largely covered by a grant of £300 from Trinity College. In itself the work probably cannot be regarded as a significant influence. Many, if not most, of the topics covered in the *Lectures* were previously published by Hamilton in journal articles. Though the *Lectures* did go beyond these publications, the fact that the subject matter was not regarded as new may help to explain why no special note was taken of it in the literature when it first appeared. Hamilton realized that the *Lectures* were, in spite of his original intentions, not suitable as an introduction for the beginner and that a new plan was called for. We know that one of England's most renowned scientists of the time, John Herschel, in spite of repeated efforts to make his way through it, only managed the first three Lectures [Hankins, 1980, 359–360].

Hamilton thus started on the *Elements of quaternions*, which grew as he worked on it from a small manual to a tome of over 800 pages when it finally came to print after his death, thanks to his son William Edwin Hamilton. A second edition, with notes and appendices by his colleague C.J. Joly, appeared in two volumes in 1899 and 1901. The *Lectures* was also supplanted by P.G. Tait's *Elementary treatise on quaternions* in 1867. Tait (1831–1901) was educated in mathematics at Cambridge University and, though his main interest was in physics, became Hamilton's closest follower and advocate. His treatise appeared in two further editions and was translated into French and German. An even more elementary *Introduction to quaternions* appeared in 1873 as a joint work with P. Kelland and went through several editions. In 1905 Joly felt there was a need for *A manual of*

quaternions wherein he reluctantly 'abandoned Hamilton's methods of establishing the laws of quaternions' while diplomatically recognizing that Hamilton's *Lectures* 'have a charm all their own' [Joly, 1905, v]. The *Lectures* thus stand less as an influence than as an historically important record of Hamilton's intentions at a key stage of development of his quaternions.

5 QUATERNIONS VERSUS VECTORS: J.W. GIBBS AND E.B. WILSON

A major impetus to the propagation of quaternions came from James Clerk Maxwell's use of them in his *Treatise on electricity and magnetism* (1873) (§44). It was from reading Maxwell that the British scientist Oliver Heaviside (1850–1925) and the U.S. mathematical physicist Josiah Willard Gibbs (1839–1903) of Yale University came independently to critically study quaternions and to develop an alternate system, vector analysis. Gibbs was also influenced by Grassmann's calculus of extension, first published in 1844 [Gibbs, 1891]. His lithographed pamphlet *Elements of vector analysis* (1881–1884) was privately printed but received rather wide circulation, even abroad in Europe. Heaviside's work, initially published in the journal *Electrician* in 1882 and 1883, was less well known (compare §49), and Gibbs became the main target of the quaternion supporters. Their theme was set by Tait who, in 1890 in the Preface to the third edition of his *Elementary treatise,* stated that 'Gibbs must be ranked as one of the retarders of Quaternion progress, in virtue of his pamphlet on *Vector analysis*; a sort of hermaphrodite monster, compounded of the notations of Hamilton and Grassmann'.

The controversy between the vector and quaternion camps is unusual in the history of mathematics in its intensity and international scope, comparable to the dispute between the followers of Isaac Newton and G.W. Leibniz over the origins and best form of the calculus. In addition to many publications from both sides, quaternionists founded an International Association for Promoting the Study of Quaternions and Allied Systems of Mathematics, which published bulletins between 1900 and 1913 [Crowe, 1967]. It should be noted that the dispute was not over the crediting of discoveries; Gibbs and other vector adherents claimed only to have a better way of achieving the same useful applications. In particular they noted that the functional usefulness of many of Hamilton's operators, such as the scalar and vector operators, S and V, could be obtained more easily without introducing the quaternion.

The modest size (83 pages) and compact style of writing in Gibbs's work stand in contrast to Hamilton's overwhelming prolixity. Furthermore, Gibbs never took the time to develop his pamphlet into a textbook in spite of the increasing popularity of his system. Instead this task fell to a former student at Yale, Edwin Bidwell Wilson (1879–1964). Wilson had studied quaternions as an undergraduate at Harvard University under J.M. Peirce. In building upon what were in effect Gibbs's lecture notes, Wilson also drew upon other works, including Heaviside's, to produce a book of 436 pages that set the pattern, with respect to notation and use, for virtually all subsequent works in vector analysis. His *Vector analysis* appeared, with a preface by Gibbs, in 1901 and was soon followed by several further printings. Initially published by Scribner's in New York, after Yale University Press was founded in 1908 it produced the second edition in 1909, incorporating corrections, and

several subsequent printings up to the Dover reprint of 1960. In spite of this substantial enlargement over Gibbs's booklet, Wilson's work was still designed as a textbook and not a treatise on the state of the subject. Each of its seven chapters included exercises. A comparison with the content of Hamilton's *Lectures* can only be made indirectly since Wilson's principal source, Gibbs, appears to have been informed about quaternions mainly through Tait's *Treatise* [Crowe, 1967, 155–158].

Wilson opens with definitions of vector and scalar quantities and the basic operations between them, paying careful attention to naming and symbolizing conventions. He uses, for example, bold letters (or 'Clarendon type' as he terms it) for a vector and ordinary type for the same letter for its scalar magnitude. (Heaviside had employed this practice and name earlier: see §49.2) Three mutually perpendicular unit vectors, $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$ and vectors as linear combinations of these are introduced. Thus a connection to the Cartesian rectangular coordinate system is immediately established.

In Chapter II the direct and skew products of vectors appear, written $\mathbf{A} \cdot \mathbf{B}$ and $\mathbf{A} \times \mathbf{B}$ for vectors $\mathbf{A}$ and $\mathbf{B}$, which have since taken on the names of dot and cross product respectively. Their correspondence to Hamilton's scalar and vector components of the product of two quaternions a and b, Sab and Vab, would have been obvious to a reader versed in quaternions. Chapters III and IV deal with the differential and integral calculus of vectors, and define the notions of derivative, divergence, curl, and scalar and vector potentials. Hamilton is credited with the introduction of the ∇ symbol for derivative—one of the few passages where Hamilton's work is explicitly mentioned.

Linear functions of vectors are the subject of Chapter V. A key concept is the 'dyad' defined as a juxtaposition of two vectors, as in $\mathbf{ab}$. Taking the dot product on the right with a vector $\mathbf{r}$ produces another vector $\mathbf{r}' = \mathbf{ab} \cdot \mathbf{r}$ that is, in this example, the product of a vector $\mathbf{a}$ and a scalar. The dyad plays a key role in the remaining two chapters which concern applications in mathematical physics and geometry, the main motivation for the subject as far as Gibbs and Wilson were concerned. There is a Section on the propagation of light in crystals that may have helped make a link to the ongoing discussions resulting from the Michelson–Morley experiments in the United States on the nature of the aether. Rotations and strains are represented by dyadic expressions (i.e. linear combinations of dyads). In particular, a dyadic reducible to the form $\mathbf{i}'\mathbf{i} + \mathbf{j}'\mathbf{j} + \mathbf{k}'\mathbf{k}$, where each of the triples $\mathbf{i}'$, $\mathbf{j}'$, $\mathbf{k}'$ and $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$ are right-handed rectangular systems of unit vectors, represents a rotation and is called a 'versor'. Here and elsewhere Hamilton's terminology is echoed. One Section of the last chapter is devoted to the representation of quadric surfaces by means of dyadics. Another Section, on curvature of surfaces, is exceptional in that it makes more use of pure vectors than of dyadics. Wilson included all the topics covered by Gibbs except applications to crystallography and the theory of orbits, topics to which Gibbs devoted much attention. Nevertheless, as he describes in his reminiscences of Gibbs [Wilson, 1931], they had virtually no interaction regarding the preparation of the book.

Of the four reviews of Wilson's work cited in [Crowe, 1967, 229], only one was unfavorable, claiming that the work should have been quaternionic. That reviewer asserted that the dyad's strong operational resemblance to the quaternion undermined any claim that the 'new' methods were really new. Furthermore, the dyad lacked the 'geometric significance' of the quaternion [Knott, 1902]. Most readers appear to have understood that indeed very little of this was essentially new. However, dyads were soon encompassed in the theory of

matrices while the role of quaternions in mathematics evolved into something somewhat less grand than Hamilton envisioned.

BIBLIOGRAPHY

Crowe, M.J. 1967. *A history of vector analysis*: *The evolution of the idea of a vectorial system*, Notre Dame: University of Notre Dame Press. [Rev. repr. New York: Dover, 1985.]

Gibbs, J.W. *Papers*. *The scientific papers of J. Willard Gibbs*, 2 vols., London: Longmans, Green. [Repr. New York: Dover, 1961.]

Gibbs, J.W. 1881–1884. *Elements of vector analysis arranged for the use of students in physics*, New Haven, Connecticut: privately printed by Tuttle, Morehouse & Taylor. [Repr. in [Gibbs, *Papers*], vol. 2, 17–90.]

Gibbs, J.W. 1891. 'Quaternions and the "Ausdehnungslehre" ', *Nature, 44*, 79–82. [Repr. in [Gibbs, *Papers*], vol. 2, 161–168.]

Graves, R.P. 1882–1891. *Life of Sir William Rowan Hamilton, including selections from his poems, correspondence, and miscellaneous writings*, 3 vols., Dublin: Hodges, Figgis; London: Longmans, Green.

Hamilton, W.R. *Papers*. *The mathematical papers of Sir William Rowan Hamilton*, 4 vols. (ed. various), Cambridge: Cambridge University Press, 1931–2000.

Hamilton, W.R. 1866. *Elements of quaternions* (ed. W.E. Hamilton), London: Longmans, Green. [2nd ed. (ed. C.J. Joly), 2 vols., 1899–1901; repr. New York: Chelsea, 1969.]

Hankins, T.L. 1980. *Sir William Rowan Hamilton*, Baltimore and London: The Johns Hopkins University Press.

Joly, C.J. 1905. *A manual of quaternions*, London and New York: Macmillan.

Knott, C.G. 1902. Review of [Wilson, 1901], *Philosophical magazine, (6) 4*, 614–622.

Pickering, A. 1995. *The mangle of practice*: *time, agency, and science*, Chicago: University of Chicago Press.

Tait, P.G. 1867. *Elementary treatise on quaternions*, Oxford: Clarendon Press.

Wilson, E.B. 1901. *Vector analysis*; *a text-book for the use of students of mathematics and physics, founded upon the lectures of J. Willard Gibbs*, New York: Scribner's; London: Edward Arnold. [Repr. New York: Dover, 1960.]

Wilson, E.B. 1931. 'Reminiscences of Gibbs by a student and colleague', *The scientific monthly, 32*, 210–227.

CHAPTER 36

GEORGE BOOLE, *AN INVESTIGATION OF THE LAWS OF THOUGHT ON WHICH ARE FOUNDED THE MATHEMATICAL THEORY OF LOGIC AND PROBABILITIES* (1854)

I. Grattan-Guinness

In this book Boole's algebra of logic received its definitive form. Influence was slow to develop, and then some changes in the algebra were made by others; but the theory became part of the fabric of logic, and also of computing in modern times. The book also contains a notable contribution to probability theory.

First publication. Cambridge: Walton and Maberly; London: Macmillan, 1854. [ix] + 424 pages.

Manuscript. Boole Papers, Royal Society Archives, London [ch. 22 missing].

Photoreprints. New York: Dover, 1958. Amherst, New York; Prometheus Books, 2003 (introd. by J. Corcoran).

Reprint. *Collected logical works*, vol. 2 (and only; ed. P.E.B. Jourdain), Chicago: Open Court, 1916. [Repr. 1940 and 1952.]

Italian translation. Indagine sulle legge del pensiero su cui sono fondata le teorie matematiche della logica e della probabilità (trans. M. Trinchero), Turin: Einaudi, 1976.

French translation. *Les lois de la pensée* (trans. S.B. Diagne), Paris: Vrin, 1994.

Related articles: Laplace on probability (§24), Grassmann (§32), Whitehead and Russell (§61), Kolmogorov (§75).

1 A SELF-MADE MATHEMATICIAN

The range and depth of the achievements of George Boole (1815–1864) are especially remarkable when one notes not only the shortness of his life but also the disadvantageous

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

circumstances of his background. He was born to an intelligent tradesman who however was so poor that George had to become the main breadwinner in his 20th year when he opened a school. Nevertheless, he found time to teach himself advanced mathematics, and also Greek, Latin, French and German, especially in order to read important works. His research papers began to appear in the early 1840s, and his principal interest soon turned to an English specialty: the 'calculus of operations', now called 'differential operators', where differentiation was represented by the letter 'D', higher-order differentiation by '$D^2, D^3, \ldots$', integration by 'D^{-1}', and so on. This tradition had developed under the influence of the algebraised calculus propounded by J.L. Lagrange (§19), initially by some French mathematicians; but from the 1810s this algebra and related topics were prosecuted in England by Charles Babbage and John Herschel as part of the revival of research mathematics there. Boole was to become a major figure in this movement in the next generation; as we see, it was to affect his work on logic.

Boole's interest in logic was very unusual for a mathematician, but it grew out of a strong renaissance on the subject that had suddenly started in 1826 with the publication of Richard Whately's *Elements of logic*. The many reactions led him to produce four more editions in a decade; among innovations was an important extension of syllogistic logic called 'quantification of the predicate', due to George Bentham in 1827 but the subject of a priority dispute during the early 1840s between the Scottish philosopher William Hamilton and Augustus de Morgan (1806–1871), the only other mathematician apart from Boole to mathematicise logic (section 7). Their non-discussion stimulated Boole to write up his first account of his theory, in the short book *A mathematical analysis of logic* ([Boole, 1847], hereafter, '*MAL*').

Through these years Boole continued with his school-teaching; but a chance for advancement came in the mid 1840s when the Queen's University of Ireland was set up, with Colleges in Belfast, Cork and Galway. Despite his lack of formal qualifications Boole was appointed Professor of Mathematics at Cork; after a delay in the organization of the University caused by the potato famine in Ireland, he took up the post in 1849 and held it until his death in 1864. He was very isolated mathematically; in particular, for some reason he seems to have had little contact with W.R. Hamilton (not to be confused with the philosopher named above). Boole's second book on logic, the subject of this article, was written in Cork; it is cited as '*LT*'.

Among biographical sources, the most significant is [MacHale, 1985]; [Diagne, 1989] may also be consulted. Some interesting obituaries were published; two are reprinted in the collection [Gasser, 2000] of old and new writings on his life and logic. See also the partial edition [Boole, 1952] of his papers and manuscripts on logic and probability theory. Recently an edition of his manuscripts on logic has appeared [Boole, *Manuscripts*]: it also contains an extensive primary and secondary bibliography for Boole, a property evident likewise in the collection [Agazzi and Vassallo, 1998], which is largely concerned with philosophical aspects of Boole's mathematics. The mathematical sides of Boole's logic are reviewed historically in [Panteki, 1992, chs. 5–8] and in somewhat modernised terms in [Hailperin, 1986, pt. 1].

2 BOOLE'S INITIAL 'ANALYSIS' OF LOGIC

The priority dispute triggered Boole to write his first book; but its content was much influenced by his researches on differential operators. Partly drawing upon his work of his friend the Cambridge mathematician Duncan Gregory, he produced a long paper on these methods which he submitted to the Royal Society in 1844. After wondering about rejecting the manuscript they published it [Boole, 1844], and then awarded him a Gold Medal for his achievement!

This theory was one of the early algebras in which the 'objects' were neither numbers nor geometrical magnitudes; and it had met controversy in its algebraic laws, such as identifying powers with orders (that is, D^2 for D on D, not D times D). Aware of the mystery, Boole (and before him Gregory) tried to bring light by highlighting the principal desirable properties of functions π and r of D on functions $q, r, \ldots$, and using them in solving various differential equations.

Three years later *MAL* appeared. In it Boole offered a novel approach to logic in the form of an algebra of the mental act x of choosing some property, and complementarily of forming the class of objects satisfying the property. An assumed universe 1 was thereby divided into complementary parts x and $(1 - x)$, and the desired laws for x were formed as closely as possible to those for the D algebra; namely, and using his notations from both sources,

1844 paper	Name	*MAL*	
$\pi rq = r\pi q$	commutative law	$xy = yx,$	(1)
$\pi(q + r) = \pi q + \pi r$	distributive law	$x(u + v) = xu + xv,$	(2)
$\pi^l \pi^m q = \pi^{l+m} q$	'index law'	$x^n = x,$	(3)

where l, m and n ($\geqslant 2$) were positive integers [1844, 225; 1847, 17–18].

The point of distinction between the two algebras was (3), where the same name was used but the laws differed; indeed $(3)_2$ was previously unknown in mathematics (apart from anticipations in Leibniz, then unpublished). The procedures in the two theories were quite distinct; the task for logic was to cast propositions into algebraic form, and then to deduce logical consequents, usually in the form of relating one of he properties (x, say) as a function of the others. To this end Boole found several remarkable expansion and elimination theorems. We shall describe them and the above laws from their appearance in his second book, the subject of this article.

3 BOOLE'S MATURE 'INVESTIGATION' OF LOGIC

LT seems to have been completed at Cork in 1852; after printing in Dublin during the following year, it appeared early in 1854 from his Cambridge and London publishers. His opening act was to stress differences from *MAL*, especially for the 'more general' methods and wider remit; here both similarities and differences will be mentioned. Its contents are summarised in Table 1.

In both books Boole showed that he saw his algebraised logic as *applied* mathematics. The title of *MAL* shows it already; and in the main text of *LT* he stated his aim as not only

Table 1. Contents by chapters of Boole's book.

Chs.	Page	Contents
1	1	'Nature and design of the work'; logic and probability theory.
2–3	24	Laws of logic; interpretations as mathematical acts or as classes; basic properties.
4	52	'Division of propositions' into 'primary' and 'secondary' (asserting truth).
5–8	66	Methods of deduction: expansion theorems, 'interpretation', elimination.
9	130	Simplification of methods, and short cuts.
10–12	150	Methods for secondary propositions.
13–14	185	Selected passages for logical analysis.
15	226	'On the Aristotelian logic'.
16–18	243	'Theory of probabilities'; methods and examples.
19	295	'Of statistical conditions' (linear programming).
20–21	320	Examples concerning causes and judgments.
22	399	'Constitution of the intellect'. [End 424.]

'to investigate the fundamental laws of those operations of the mind by which reasoning is performed' but also 'to give expression to them in the symbolic language of a Calculus' (p. 1). It came clear that the proposed theory was normative; he did not try to treat the mysterious depths of *actual* thought.

The basic properties of mental operations were laid out in chs. 2 and 3 although, in a change from *MAL*, the letters $x, y, \ldots$ were taken much more often to refer to classes than to mental acts. Throughout he took classes in the traditional part-whole sense, not within the set theory of Georg Cantor, still decades away (§46); further, the phrase 'let x represent "all men," or the class "men."' (p. 28) should not be over-interpreted as anticipating quantification theory.

Among the manners of combination, addition was defined only between disjoint classes, giving the clause 'let $+$ stand for "*and* " and "*or*"' (p. 33)—odd to read in a logic book. Although Boole did not say so, he might have wished to avoid defining multi-classes, where the members of the overlap were counted twice. At all events, his restriction of addition was not well received, as we shall see in section 7.

Boole symbolized intersection by concatenation 'xy', as in algebraic multiplication, and the laws (1)–(3) were duly emphasized. He defined complementation '$x - y$' when y was a sub-class of x, and associated with 'except' (pp. 29–34). He assumed various—or maybe overlooked—further laws for these operations, especially associativity; and he did not have a symbol for negation, so that he could not properly express the proof-method by contradiction.

The over-arching class was now called 'universe of discourse' (p. 42); and an improvement over *MAL* was that it was not taken to be absolutely universal, for then truth by content merges with truth by form (for example, 'London is a city' and 'London is a city or London is not a city' are both true, but for different reasons). Such a restriction had been

imposed by logicians for centuries in some way or other; Boole made it canonical here. He symbolized his universe as '1', and its complementary class as '0'; but he fell into an unclear aspect of part-whole theory in calling it 'Nothing' but not really differentiating it from no thing (pp. 47–48). However, it allowed him to derive from $(3)_2$ the '*principle of contradiction'*, also to be called 'the law of duality' (pp. 49–51):

$$x(1-x) = 0;\ \text{curiously},\ x + (1-x) = 1 \tag{4}$$

was not presented despite its equally ancient linkage to the law of excluded middle (compare p. 76).

In ch. 3 Boole also allied his basic laws and the consequent properties with 'the laws of the operations of the human mind', partly rehearsing the material again in the concluding ch. 22 'On the constitution of the intellect'. With regard to language, he associated classes with nouns and adjectives: 'good, good men = good men', 'white men = men white' (pp. 32, 44) and so on. Like most logicians until modern times, he did not explore the logic of adverbs.

Next Boole introduced a 'division of propositions' into 'primary' ones concerned with things, and 'secondary' ones which dealt with propositions; for example (his), 'the sun shines' and 'it is true that the sun shines' (pp. 52–53). To us this seems to be the distinction between a language and its metalanguage, but this was to be fully grasped only 80 years later (see §71 on Gödel). Further, in a rather ungainly way he treated hypothetical compound propositions as secondary (p. 53). A proposition X was grasped by 'an act of mind' x to be true for some period x of time (p. 165: note two uses for 'x'); if X were true all the time, then '$x = 1$', and if false, never ('$x = 0$'). The basic laws and properties were maintained (chs. 11–12); for example, xy now denoted the time during which both X and Y were true simultaneously. While the algebra is unexceptionable, the philosophical status of time in a theory of mental acts seems problematic.

A major difference from *MAL* was the status of syllogistic logic. In the earlier book it had provided many of the examples analysed; here it débuts only on p. 226, in the last of the chapters on logic. Boole had realized in the meantime that he had advanced his algebra of logic far beyond the Aristotelian tradition.

4 THE ALGEBRAIC METHODS OF DEDUCTION AND ELIMINATION

Given a collection of logical premises, one main aim of Boole's logic was to choose one of the given properties as subject and find, as the logical deduction, its relationship to the other properties. The algebra fulfilled this aim by means of expansion theorems and algebraic means of eliminating properties (ch. 5). For example, for a function '$\phi(xy)$' of two mental acts (or of classes) the expansion was

$$\text{`}\phi(xy) = \phi(00)(1-x)(1-y) + \phi(01)y(1-x) + \phi(10)x(1-y) + \phi(11)xy\text{'};\tag{5}$$

the coefficients were calculated simply by giving the values 0 or 1 as appropriate to x and y. (5) was a linear expansion, which Boole expressed generally as '$a_1t_1 + a_2t_2 +$ &c.' (p. 93).

The function could take various values, with four outcomes for its coefficients (ch. 6, where he used only the classical interpretation of the letters). For the value 1, set the associated term $= 0$, and offer the resulting equation as one of the deductions); for 0, drop the term; $0/0$ gives the indeterminate portion of the solution, where one adjoins to the associated term (u, say) the indeterminate class 'v' and read 'uv' as the overlap); finally, for and any other value, including $1/0$, put the term $= 0$ and present that proposition as a condition under which the solution obtains).

For example, a proposition about types of beasts came out as

$$x = yz;\ \text{thus, as subject,}\ 1 - y = (z - x)/z. \tag{6}$$

The expansion of $(6)_2$ gave as solution

$$1 - y = (1 - x)z + v(1 - x)(1 - z), \quad \text{with } x(1 - z) = 0 \tag{7}$$

as side condition, because the coefficients of the terms were respectively 1, $0/0$ and $-1/0$; in addition xz took the coefficient 0 and so disappeared (p. 94).

Boole did not notice that some of his deductions also admitted singular solutions involving, for example, the empty class [Corcoran and Wood, 1980]. This is disappointing, since he emphasized such solutions in his work on differential equations, and they exemplified his basic law (3) of logic; on the one hand (like an x) they were solutions but on the other hand (like a $1 - x$) not part of the general solution.

Expansion (5) had already been presented in *MAL*: new in *LT* was a theory of 'elimination' (ch. 7) based upon the theorem that any logical function $f(x)$ satisfied the equation

$$f(0)\,f(1) = 0; \quad \text{for two variables '}\phi(1,1)\phi(1,0)\phi(0,1)\phi(0,0) = 0', \tag{8}$$

and so on. Thereby x (or x and y) were removed from the deduction, leaving a relationship between the other variables. One of his examples (p. 105) removed the indeterminate class v by treating the proposition

$$y - v(1 - x) = 0 \text{ as a function of } v;\ \text{then } (7)_1 \text{ gave } yx = 0. \tag{9}$$

However, elimination of v from $vx = vy$ leads only to $0 = 0$. Then followed an account of 'reduction' (ch. 8), where a collection of n equations $V_r = 0$ was handled by conversion to the single equation

$$\sum_r a_r V_r = 0,\ a_r \text{ arbitrary constants, or to } \sum V_r^2 = 0. \tag{10}$$

A novel feature of Boole's algebra was his use of *interpretation*: the equation(s) expressing the premises should be interpretable, and the logical consequence also; but the intervening lines of deduction need not take an interpretation, nor need the values of the coefficients be other than 0, v or 1; hence the role of $x(1 - z)$ in the side condition $(7)_2$.

5 BOOLE'S TREATMENT OF PROBABILITY THEORY

The major difference between *MAL* and *LT* was the appearance in the latter of probability theory; it took up chs. 16–21, at over 150 pages. Around 1849 he had realized that compound events could be handled in his logic as con- and/or disjunctions of simple ones, and so their consequences determined by his laws and expansions theorems and any attendant probabilities calculated accordingly. By these means he hoped to bring a new level of generality to the theory (p. 265) with probability logic [Hailperin, 1986, pt. 2].

Boole's construal of probability was epistemic: 'the word *probability*, in its mathematical acceptation, has reference to the state of our knowledge under which an event may happen or fail. [. . .] Probability is expectation founded upon partial knowledge' (p. 244). However, he did not always distinguish the probability of a conditional proposition from conditional probability.

In a remarkable chapter 'on statistical conditions' Boole considered situations in which the values of some or all probabilities may be known only approximately, or to within some upper and/or lower bounds. In his first case he showed that the probability of a disjunction of events was less than the sum of the probabilities of each event (pp. 297–299), an inequality now named after him as part of his modest influence on the subject. Some of the more elaborate later cases led him to aspects of linear programming, which was not to develop as a mathematical topic for nearly a century although his was not the first anticipation [Grattan-Guinness, 1994]. A few scientific examples and case studies were tackled, but no religious ones, not even buried in sentences. This last silence merits discussion.

6 THE RELIGIOUS CONNOTATION OF BOOLE'S LOGIC

For Boole an important aspect of his logic was its connection with religion. During his adult career British Christianity was in a state of considerable ferment, with the strong rise of Dissenting versions competing with each other and with the established Church of England. Boole belonged to one of these factions: ecumenism, which advocated the One and Only God in contrast to establishment Trinitarianism. This stance was reflected in his logic by the status of the universe 1, to be divided into its components. The link was exhibited in *LT*, though without announcement and so overlooked by most readers. The clearest evidence is provided in ch. 13, where he provided logical analyses of propositions due to Samuel Clarke and Benedict Spinoza concerning the necessary existence of '*Some one unchangeable and independent Being'* (p. 192). He also alluded to his position in print very discreetly a few lines from the end of the book, where he referred to 'those who profess an intellectual allegiance to the Father of Lights', one of the standard Dissenter names of the Godhead (not God as Orthodoxedly construed). He greatly admired the book *Philosophie— Logique* (1855) of Father A. Gratry, who larded his own version of logic with religious fervour.

To Boole, and also to his wife Mary (whom he married in 1855) the hero was Frederick Denison Maurice (1805–1872), who advocated ecumenism with great force in mid century and so was dismissed from his chair of Divinity at King's College London. Boole spent parts of several summer vacations in England in his last years, studying in London

libraries—and on the Sundays he attended Maurice's services. The strength of his admiration was exhibited in his last days. Late in November 1864 he walked to the University in the rain without protection, and after lecturing in wet clothes he soon developed pneumonia. As he lay at home on his deathbed, he asked that a portrait of Maurice be set up alongside.

This interpretation of Boole's logic was influential only upon his widow, who continued to prosecute it and especially the educational aspects of his philosophy after his death. In penury after his death with five young daughters to bring up, she obtained employment for some years from Maurice in Queen's College, an establishment for female education that he had established in London.

7 BOOLE'S GRADUAL INFLUENCE

After publishing LT Boole put out a few papers on probability theory, but not on logic. However, he attempted a more general book on the subject, and also one on the philosophy of mathematics; neither was finished, but the major manuscripts have recently been published, along with some others from earlier periods [Boole, *Manuscripts*]. No radical revision of theory emerges from these sources, but in his planned book on logic he elaborated on the relationships between logic and reasoning: interestingly, the religious side was not rehearsed.

After 1854 Boole largely went back to the differential and integral calculus, producing successful textbooks in differential and difference equations (1859 and 1860, and later editions); his wife helped to check the accuracy of the solutions to exercises. The books related in part to his teaching at Cork (in contrast to logic, which he seems never to have taught). Indeed, this side of his research work was far better appreciated than his logic, which was regarded as an interesting curiosity but marginal to mathematicians' concerns. Further, of the two books *MAL* gained *more* of the modest attention. This was the opinion of de Morgan, the other major British figure of the time working on the algebraisation of logic, in his case on the symbolization of various aspects of syllogistic logic and especially its extension with a logic of relations [Merrill, 1990]. He and Boole had quite a lengthy correspondence [Smith, 1982], but on matters logical they tended to talk past each other [Corcoran, 1986].

Gradually LT picked up a public. In particular, it was the text used by the first serious reader of Boole's logic, the mathematician and economist Stanley Jevons (1835–1882). While broadly happy with Boole's theory, especially with the new laws for logic, he disliked expansion theorems and especially the restriction of union to disjoint classes. They conducted a (non-)correspondence in 1863 and 1864, shortly before Boole's death [Grattan-Guinness, 1991]. Jevons concentrated upon the meaning of '$+$' and the class expression '$(x + x)$': for Jevons it equalled x, while for Boole it was not interpretable, although his expansion theorems showed that the (interpreted) equation $x + x = 0$ took the solution $x = 0$. In another change, Boole's temporal theory of secondary propositions was in effect replaced by the propositional calculus, in 1877 by Hugh MacColl, and two years later by Gottlob Frege.

Boole's most loyal follower was John Venn, mainly in his book *Symbolic logic* (1881, 1894: the origin of this name, by the way); however, even this Cambridge Reverend did

not adopt the religious connotations (nor did he adopt Boole's reading of probability theory). With other logicians Jevons's view prevailed, especially when the American logician C.S. Peirce came to the same conclusion soon afterwards, in a combination of a modified Boolean system and an elaboration of De Morgan's logic of relations. This theory was extended by the German mathematician Ernst Schröder, in a vast collection of *Vorlesungen über die Logik der Algebra* (1890–1905). A few other mathematicians and philosophers took interest; for example, in Russia [Styazhkin, 1969, ch. 6]. Sometimes the interest was furthered in conjunction with reactions to the algebra of Hermann Grassmann ([Peckhaus, 1997, ch. 6]; and §32). The 1916 edition of *LT* by P.E.B. Jourdain (who added a few notes and an index) made the book more available.

But after and even during Schröder's mammoth efforts algebraic logic rather floundered, becoming eclipsed within symbolic logic itself by the mathematical logic developed by Frege and especially by Giuseppe Peano and his followers A.N. Whitehead and Bertrand Russell ([Grattan-Guinness, 2000, esp. chs. 2–7]; see also §61). The name 'Boolean algebras' was introduced by the American logician H.M. Sheffer in 1913, referring to them just as algebras as such. They had become part of the furniture of logic, especially for the propositional calculus: the applications to electrical circuit theory, to communication and computing, and to neurophysiology started only from the late 1930s onwards.

And here lies a great irony. Part of Boole's philosophy was the claim that the mind can grasp the general from a few instances of the particular, rather than accumulate the general only from accumulations of individual cases, the view then advocated especially by J.S. Mill. Boole seems to have formed his position during (and maybe because of) his early experience as a teacher; he never took an interest in Charles Babbage's long saga to create a mechanical computing machine, for it was based merely on repetition. Now Babbage's work presaged the modern computer, which is the chief source of the ubiquity of Boole's name today!

BIBLIOGRAPHY

Agazzi, E. and Vassallo, N. (eds.) 1998. *George Boole. Filosofia, logica, matematica*, Milan: FrancoAngeli.

Boole, G. 1844. 'On a general method in analysis', *Philosophical transactions of the Royal Society of London, 134*, 225–282.

Boole, G. 1847. *The mathematical analysis of logic*, Cambridge: Macmillan; London: Bell. [Repr. 1948, Oxford: Blackwell; in [1952], 49–124; and in P. Ewald (ed.), *From Kant to Hilbert*, vol. 1, Oxford: Clarendon Press, 1996, 451–509. French trans.: in F. Gillot, *Algèbre et logique*...Paris: Masson, 1962, 13–88. Italian trans. Turin: Boringheri 1965. German trans. Halle/Saale: Hallescher Verlag, 2002.]

Boole, G. 1952. *Studies in logic and probability* (ed. R. Rhees), London and La Salle, Ill.: Open Court.

Boole, G. *Manuscripts. Selected manuscripts on logic and its philosophy* (eds. I. Grattan-Guinness and G. Bornet), 1997, Basel: Birkhäuser.

Corcoran, J. 1986. 'Correspondence without communication', *History and philosophy of logic, 7*, 65–75.

Corcoran, J. and Wood, S. 1980. 'Boole's criteria of validity and invalidity', *Notre Dame journal of formal logic, 21*, 609–638. [Repr. in [Gasser, 2000], 101–128.]

Diagne, S.B. 1989. *Boole. L'oiseau de nuit en plein jour*, Paris: Belin.
Gasser, J. (ed.) 2000. *A Boole anthology*, Dordrecht: Kluwer.
Grattan-Guinness, I. 1991. 'The correspondence between George Boole and Stanley Jevons, 1863–1864', *History and philosophy of logic, 12*, 15–35.
Grattan-Guinness, I. 1994. ' "A new type of question": on the prehistory of linear and non-linear programming, 1770–1940', in E. Knobloch and D. Rowe (eds.), *History of modern mathematics*, vol. 3, New York: Academic Press, 43–89.
Grattan-Guinness, I. 2000. *The search for mathematical roots, 1870–1940. Logics, set theories and the foundations of mathematics from Cantor through Russell to Gödel*, Princeton: Princeton University Press.
Hailperin, T. 1986. *Boole's logic and probability*, 2nd ed., Amsterdam: North-Holland.
MacHale, D. 1985. *George Boole—his life and work*, Dublin: Boole Press.
Merrill, D.D. 1990. *Augustus De Morgan and the logic of relations*, Dordrecht: Kluwer.
Panteki, M. 1992. 'Relationships between algebra, differential equations and logic in England: 1800–1860', Ph. D., C.N.A.A. (London).
Peckhaus, V. 1997. *Logik, Mathesis universalis und allgemeine Wissenschaft. Leibniz und die Wiederentdeckung der formalen Logik im 19. Jahrhundert*, Berlin: Akademie-Verlag.
Smith, G.C. (ed.) 1982. *The Boole–De Morgan correspondence*, 1982, Oxford: Clarendon Press.
Styazhkin, N.I. 1969. *From Leibniz to Peano: a concise history of mathematical logic*, Cambridge, MA: MIT Press.

CHAPTER 37

JOHANN PETER GUSTAV LEJEUNE-DIRICHLET, *VORLESUNGEN ÜBER ZAHLENTHEORIE*, FIRST EDITION (1863)

Catherine Goldstein

The *Vorlesungen*, based upon Lejeune-Dirichlet's lectures delivered in 1856–1857 in Göttingen, were posthumously published by Richard Dedekind, who enriched them with substantial supplements, incorporating material both from Dirichlet and from himself. This authoritative and carefully written introduction played a decisive role in attracting mathematicians of the second half of the 19th century to number theory. It provided a bridge between C.F. Gauss's *Disquisitiones arithmeticae* (1801) and the development of the theory of algebraic number fields as promoted by David Hilbert in his 1897 *Zahlbericht*.

First edition. (Ed. Richard Dedekind), Braunschweig: Vieweg, 1863. xiii + 414 pages. [Also available in electronic form on Gallica, http://gallica.bnf.fr.]

Later editions. 2nd 1870–1871, 3rd 1879–1880, 4th 1894 (photorepr. New York: Chelsea, 1968). All revised and augmented by Dedekind, and same publisher.

Italian translation. Lezioni sulla teoria dei numeri (trans. Aureliano Faifofer), Venice: Tipografia Emiliana, 1881–1882.

Partial Russian translation. Teoryia chisel (trans. J.M. Nasarjevsky), Saint Petersburg: 1899. [First three sections of the *Vorlesungen*, supplemented with exercises.]

English translation of the 1st edition. Lectures on number theory (introd. and trans. John Stillwell), Providence: American Mathematical Society; London: London Mathematical Society, 1999. [Without Dedekind's preface; some footnotes and details taken from later editions.]

French translation. Announced by Edouard Lucas in his *Notice sur les travaux scientifiques* (1880), with the collaboration of M. Tastavin, but apparently never published.

Related articles: Gauss on number theory (§22), Dedekind on arithmetic (§47), Weber (§53), Hilbert on number theory (§54).

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

1 A POSTHUMOUS TEXTBOOK

Johann Peter Gustav Lejeune Dirichlet (1805–1859) played a decisive role in propelling German mathematics to the forefront of European science. Born in Düren, he went as early as 1822 to Paris, then the center of mathematical research, where he followed lectures at the *Collège de France* and the *Sorbonne*. His early acquaintance with the analytic works of French mathematicians, such as Joseph Fourier, constitutes one of the two main components of his formation (compare §39.2). The other one is a life-long and deep involvement with *Disquisitiones arithmeticae* by C.F. Gauss (1777–1855), which provided inspiration to him both for its mathematical themes and rigorous reasoning [Gauss, 1801] (§22). Having been made aware of Dirichlet's work through Alexander von Humboldt, in 1826 Gauss mentioned a small memoir by Dirichlet on higher arithmetic to the secretary of the mathematical section at the Berlin Academy of Sciences, Johann Encke. Gauss wrote that it revealed an excellent talent [Gauss, *Works*, vol. 12, 70]:

> The phenomenon pleases me all the more as examples of somebody who is acquainted with these topics are rare—I know almost none in Germany—and as I am convinced that this is also one the best means to sharpen mathematical talent for other, very different, branches of mathematics. And it would be all the more distressing if his homeland, Prussia, lets itself be outstripped and if France were to appropriate for itself this excellent talent.

Such recommendations provided Dirichlet with a position in Breslau (1827), then in Berlin (1828–1855), at the Military School and the University, before succeeding Gauss at the University of Göttingen in 1855. Sixty years after Dirichlet's death, Felix Klein still emphasized the lasting importance of Dirichlet's lectures in shaping mathematical training in German universities and in providing a model of what a course should be [Klein, 1926–1927, 96]. They inspired and motivated a variety of mathematicians such as Gotthold Eisenstein, Leopold Kronecker, Bernhard Riemann, Paul Bachmann—and Richard Dedekind (1831–1916).

In letters and commentaries to his edition, Dedekind has described in some detail the elaboration of the *Vorlesungen.* 'At the time your father moved from Berlin to Göttingen during the autumn 1855', he explained to Walter Lejeune-Dirichlet in 1876 [Scharlau, 1981, 51–52]:

> I was already a Privatdozent, but I welcomed the fortunate opportunity and attended his lectures; I did not take any notes during them, in order to listen more carefully, and the highly penetrating presentation inspired me to write down in the shortest manner its most essential moments, at home, from memory. When I had become gradually more acquainted with your father, I showed him from time to time these exercise books: as he himself never wrote down his lectures and entertained the idea of publishing at least those on number theory, my notes were welcomed by him as giving an approximate overview of the extent of the various parts and he often discussed this draft with me.

After Dirichlet's early death in 1859 prevented him from fulfilling his project, Dedekind took over the task of publishing the winter course of 1856–1857, which, 'although be-

ginning with the elements, was chiefly devoted to the theory of quadratic forms and handled it more completely than in the preceding years' [Dedekind, 1864]. He used his 'extremely short' daily notices alluded to above, 'which contained almost only the key parts of the proofs'; but he also followed Dirichlet's wish to add several complements in order to round off the textbook to form a more satisfying unity, working assiduously on the project for about three years ([Haubrich, 1992, 158], drawing upon the correspondence between Dedekind and the publisher Vieweg).

Let us leave the floor to [Dedekind, 1864] to sketch the contents of the book:

> [T]he first section deals with divisibility, the second with the congruence of numbers, the third with quadratic residues; in the fourth the elements of the theory of binary quadratic forms are presented and the fifth contains the solution, first given by Dirichlet, of the problem of determining the number of classes in which the binary quadratic forms of given determinant are distributed. Besides the main course, properly speaking, Dirichlet had a supplementary course in which some important auxiliary results, pertaining to other fields, were proved; this separation has been preserved, in order not to interrupt the course of thought of the fifth section which would not be easy for a beginner to grasp; the content of this auxiliary course is given in the first three supplements. The following supplements (IV–IX) are additions, through which the editor has tried to round off the domain of the material handled in the sense indicated above.

These supplements mainly reproduced papers by Dirichlet or other known results.

For the subsequent editions, Dedekind made a few changes to the main text, and added explicit references to the literature. He explained these additions by his wish to 'awake in the reader an image of the progress of science, the truths of which, both deep and distinguished, form a treasure which is the imperishable fruit of an authentically noble competition among European peoples' (preface of the 1871 edition), a stance which acquires a particular resonance, written as it was in the middle of the Franco-Prussian war. Moreover, Dedekind added an important supplement in 1871, cut in two in 1879 and 1894); they treated the composition of forms, a crucial but extremely difficult topic in Gauss's book, which served here as a ground for Dedekind's own theory of algebraic numbers and ideals. The contents of the first edition are summarised in Table 1.

2 THE SIMPLIFICATION OF GAUSS'S *DISQUISITIONES ARITHMETICAE*

As Henry Smith remarked at the beginning of his report on number theory for the British Association for the Advancement of Science, Gauss's *Disquisitiones* was still in the 1850s the classical source for the theory of numbers, together with Adrien-Marie Legendre's more elementary *Théorie des nombres* (1830) [Smith, 1859–1865, 38]. However, its synthetic approach, loaded with lengthy computations, as well as the complexity of its subject matter, made it a daunting Everest for most mathematicians. Dirichlet entertained a deep relation to this book all during his life and in a number of his research papers tried to give a 'clear and appropriate elaboration' [Kummer, 1860, 332] of various aspects and to transmit the core of the book to a larger audience. In Hermann Minkowski's words [1905, 151]:

Table 1. Contents by Sections of the lectures. xiii + 414 pages.

Ch. or Suppl.	**Sections**	**Short description of the contents**
Preface		
Chapter I. Divisibility of numbers	1–2	Product of several numbers, commutativity.
	3	Divisibility.
	4–7	G.c.d. and l.c.m.
	8–10	Prime numbers, divisors.
	11–14	Euler Phi function.
	15	Divisors of $m!$
	16	'Looking back'.
Chapter II. Congruence	17–20	Congruences, residues, generalized Fermat theorem.
	21–26	Congruences with unknowns.
	27	Wilson's theorem.
	28–31	Power residues, primitive roots.
Chapter III. Quadratic residues	32–39	Quadratic residues and non-residues, Legendre symbol.
	40–41	Primes with -1 or 2 as quadratic residue.
	42–44	Reciprocity law, content and first proof.
	45–47	Jacobi symbol.
	49–51	Second proof of reciprocity law.
	52	Linear forms containing primes.
Chapter IV. Quadratic forms	53	Binary quadratic forms.
	54–56	Transformations of forms, equivalence.
	57–58	Two-sided forms.
	59–63	Division of forms into classes, representation of numbers, reductions of the problem of classification.
	64–67	Forms with negative determinants.
	68–71	Particular cases.
	72–85	Forms with positive determinants, associated roots, periods of reduced forms, Fermat–Pell equation.
Chapter V. Class number of binary quadratic forms	86	Numbers properly represented by primitive forms.
	87	Number of representations.
	88–90	Fundamental equation.
	91	Decomposition into two squares.
	92–95	Further work on the fundamental equation.
	96–97	Expression of the class number.
	98-105	Fundamental equation for positive determinant.
	106–110	Formulas for the class number.
Supplement 1	111–116	Lemmas for ch. V arising from the theory of circle division.
Supplement 2	117–119	Limiting value of some infinite series.
Supplement 3	120	Connection between area and number of lattice points.

Table 1. (*Continued*)

Ch. or Suppl.	Sections	Short description of the contents
Supplement 4	121–126	Genera of quadratic forms, representation of numbers, characters.
Supplement 5	127–131	Power residues for composite moduli.
Supplement 6	132–137	Primes in arithmetical progressions.
Supplement 7	138–140	Results from the theory of circle division.
Supplement 8	141–142	Approximation of quadratic surds and Fermat–Pell equation.
Supplement 9	143–144	Convergence and continuity of some infinite series.

> Dirichlet did not study this work once or several times only, he never stopped his entire life recalling to his mind again and again the stock of deep thoughts which it contains. Sartorius von Walthershausen said once: exactly as certain priests wander around with their prayer book, Dirichlet used to go on all his travels only in company of a much read, battered copy of the Disquisitiones Arithmeticae.

The simplifications introduced in the *Vorlesungen* with respect to the *Disquisitiones* are manifold and operate at several levels. One, for instance, is in the order of presentation: while Gauss introduced congruences from the start and deduced the greatest common divisor of two numbers from their factorization into primes [Gauss, 1801, arts. 16 and 18], Dirichlet begins with the properties of divisibility of integers, and derives the unique factorization theorem from the Euclidean algorithm (Sect. 8). Indeed, he even summarizes his first chapter by saying that 'the whole structure rests on a single foundation, namely the algorithm for finding the greatest common divisor of two numbers' (Sect.16); this way of bringing out the principles, in most cases very simple, on which proofs or theories are constructed, is quite characteristic of Dirichlet's practice and occurs also in more complicated contexts.

For instance, the eighth supplement is devoted to the study of the Fermat–Pell equation $T^2 - DU^2 = 1$, for D not a square. A key lemma for the existence of integral solutions of this equation is the proof that there are always infinitely many pairs of integers x and y such that $x^2 - Dy^2 < 1 + 2\sqrt{D}$. This lemma can be derived from the approximation of the irrational $\sqrt{D}$ arising from its expansion in continued fraction. But a simple alternative lies in another 'Dirichlet principle', the pigeonhole principle: for each m and for each integral value of y between 0 and m, it is easy to find a unique value of x such that $0 \leqslant x - y\sqrt{D} < 1$. Dividing the interval between 0 and 1 into m segments of length $1/m$, one sees that two of these couples x, y, for different y, should be in the same interval (pigeonhole principle). A solution of our inequality is thus found, whatever the m chosen, thereby providing an infinity of them.

Another type of reworking concerns the most celebrated result of the *Disquisitiones*, the law of quadratic reciprocity. It states that, for p and q any two odd primes, an integer is a quadratic residue (that is, is congruent to a square) modulo one of them if and only if it is also a quadratic residue modulo the other—except if p or q are both of the form $4n + 3$, in which case a number is a quadratic residue modulo p if and only if it is not a

quadratic residue modulo q. Complementary laws exist also for $p = 2$ and to decide modulo which numbers -1 is a quadratic residue. This law, which Gauss called 'fundamental', was proved for the first time, twice, in the *Disquisitiones*. Dirichlet chooses to present first a third, shorter, proof, also due to Gauss, and based on the so-called Euler criterion, namely the fact that a number D prime to p is a quadratic residue modulo p or not, according as $D^{(p-1)/2}$ is congruent to 1 or -1 modulo p (Sects. 43–44). He then gives Gauss's first proof based on a complete induction, but again simplifies the lengthy original discussion involving numerous cases by using a generalization of the Legendre symbol, the Jacobi symbol (Sects. 48–51).

However, the most striking effects of Dirichlet's trimming concern the theory of quadratic forms. He chooses to concentrate on the problem of equivalence of forms (considering only briefly the more general situation studied by Gauss of one form containing another), and drops completely the marginal case of forms of determinant 0. But then he sketches in a crystal clear manner what has since then become the standard steps for the study of forms: 'to decide whether two given forms of the same determinant are equivalent and hence members of the same class; to find all substitutions that send one of two equivalent forms into the other'(Sect. 59), reducing to them the classical problem of the representation of numbers by forms (Sect. 60). Two quadratic forms are said to be equivalent (or in the same class) if one can be deduced from the other by a linear transformation of the variables, with integral coefficients and determinant ± 1. A crucial ingredient in the classification of forms, at least since J.L. Lagrange, is the theory of reduction: any binary quadratic form of determinant D is equivalent to a so-called *reduced form*, with the same determinant and with coefficients satisfying simple inequalities; for a given D, these inequalities can be valid only for a finite number of forms, hence there is only a finite number of reduced forms and thus a finite number of classes of forms for a given determinant. For negative D, there is essentially exactly one reduced form per class, but this is no longer the case for positive determinants. Reduced forms for a given positive determinant can be distributed into periods of reduced forms of the same class. Dirichlet noticed that by associating to a binary quadratic form $ax^2 + 2bxy + cy^2$ of positive determinant $D = b^2 - ac$ the (complex) roots of the equation $ax^2 + 2bx + c = 0$, one can derive most of the facts about reduction from the study of these roots; in particular, from their expansion into continued fractions. It is this presentation, simpler and more evocative than the computations on the coefficients of forms used by Gauss, that Dirichlet chooses in his fourth section. He also simplifies other aspects of the theory of forms by using analytic methods, as we will see in the next section.

As Haubrich has aptly described it, Dirichlet was considered a master of proof analysis, that is the art of reflecting on proofs in order to understand and to simplify their functioning and presentation; according to C.G.J. Jacobi, '[Dirichlet] alone, not I, not Cauchy, not Gauss, knows what a complete rigorous mathematical proof is, but we know it only from him' (quoted in [Haubrich, 1992, 14, note 53]). Again, Kummer commented that, for the *Vorlesungen* as for his other arithmetical works, 'one recognizes in general that the methods through which Dirichlet has simplified number theory and made it more easily accessible, are created essentially out of a fundamental study of more general theories; the proofs of statements do not rely on special and accidental determinations, but usually on the essential

properties of the number-theoretical concepts concerned and thus communicate even in particulars a knowledge of the general' [Kummer, 1860, 333].

3 ANALYSIS AND ARITHMETIC

According to Dirichlet, 'the characteristic feature of [t]his method [to simplify the theory of forms with positive determinant] is that it brings irrational numbers into the circle of our ideas' (Sect. 72). As mentioned above, analysis and Gauss's number theory were two key features for Dirichlet's mathematical orientation and the synthesis of these two was one of his most celebrated contributions to mathematics. According to Kummer [1860, 327]:

> In his mind striving to unity everywhere, he could not let these two spheres of thought alone without exploring their internal relations, in which he looked for and indeed found the knowledge of many deeply hidden properties of numbers. His applications of analysis to number theory, which resulted from it, are distinguished from all other previous analogous attempts, mainly because in them analysis has been adopted into the service of number theory in such a way that it not only bears a few accidental isolated results but should provide necessarily the solutions of certain general classes of arithmetical problems, which are still inaccessible by other paths.

Two main applications of these ideas, dating from the end of the 1830s, are present in the *Vorlesungen*: the computation of the class number of forms for a given non-zero determinant and the theorem according to which 'each unbounded arithmetic progression $kx + m$, whose initial term m and difference k are relatively prime, contains infinitely many primes' (Sect. 137). The first is the core of the fifth chapter (with the help of the first supplements), the second fills the sixth supplement.

The point of departure for the second result—used without justification by Legendre in his attempt to prove the reciprocity law—is the equality, due to Euler, between the product $\prod_p (1 - 1/p^s)^{-1}$, taken only over primes p, and the series $\sum_n n^{-s}$, taken over all integers n, for $s > 1$. As the series diverges when s goes to 1, the product should diverge too, which means that there should exist an infinity of primes. To adapt this idea to primes belonging to arithmetical progressions with difference k, Dirichlet introduces the L-series, $L = \sum \psi(n)$, where $\psi(n)$ is a multiplicative function such that L absolutely converges: in the application to arithmetical progressions, $\psi(n)$ is defined, for n prime to k, as $n^{-s} \times$ a product of roots of unity associated with the various prime factors of k. The series can also be expressed, for $s > 1$, as a product over the primes (not dividing k), and an adequate combination of the logarithms of the various L-series equals, up to a convergent series, a series of inverse powers taken over all the primes belonging to the progression under consideration. To establish the theorem, it is thus required to prove the divergence of the total expression, and thus to study the behaviour of the various L-series near 1. When all the roots of unity are 1, the corresponding L-series is the arithmetical series, up to a finite factor, and diverges. The case where complex roots of unity appear can be settled easily once the other cases are dealt with. The most delicate case is then when the roots of unity entering the definition of the L-series are all real (that is, ± 1). If the difference k

is itself a prime number, an evaluation of the series can be made through Fourier analysis or integral calculus; but in general, Dirichlet needs to identify the L-series with one of the factors entering the expression of the number of classes of quadratic forms for a specific determinant, linked to certain divisors of k: as this class number is a non-zero integer, this L-series does not vanish. Putting together these facts for the various L-series proves the theorem.

The expression for the class numbers occupies the final chapter of the *Vorlesungen* and constitutes the main addition to Gauss's results in the core of the text. Again, it relies on the consideration of specific series, here of the type

$$\sum\sum \psi\left(ax^2 + 2bxy + cy^2\right)^s \quad \text{with functions } \psi(n) \text{ as above,} \tag{1}$$

the sum being taken over quadratic forms of a given determinant and over integral values of x and y satisfying certain conditions. The double sums are evaluated through two different groupings, one of them being such that the limit of the components when s goes to 1 does not depend upon the quadratic form considered, but only upon the determinant, which makes the class number for this determinant appear. The result, and the path followed, varies notably according as the determinant is positive or negative. For instance, if the determinant D is negative, odd, without square divisors and of the form $4n+1$, the number of classes for the corresponding quadratic forms is $h(D) = \sum_0^4 (s/|D|)$, where $(s/|D|)$ denotes the Jacobi symbol (Sect. 106). For a positive determinant D with otherwise the same properties, one finds

$$h(D)\log\left(T + U\sqrt{D}\right) = -\left(4 - 2\left(\frac{2}{D}\right)\right)\sum_0^1 \left(\frac{n}{D}\right)\log\sin\frac{n\pi}{D}, \tag{2}$$

where T and U are the smallest solutions of the Fermat–Pell equation $T^2 - DU^2 = 1$. Establishing these formulas and others uses several lemmas about the convergence of series and about integration; they are generally relegated to the various supplements.

The study of L-series and their variants (later also for a complex variable) became an important topic in number theory. In the *Vorlesungen*, besides the results just mentioned, they are also used to prove for instance that forms are equally distributed among genera (supplement IV) and other recondite questions arising from the *Disquisitiones*. As the quotes given above testify, the exploitation of such analytic techniques and the intervention of transcendental functions may appear as positive features, underlining the unity of mathematics. However, they will appear to be a drawback to the tenants of the programme of arithmetization who will, on the contrary, try to eliminate them from proofs of arithmetical results.

4 DEDEKIND'S SUPPLEMENTS X AND XI: TOWARDS THE THEORY OF IDEALS

In the second edition (1871) Dedekind added a tenth supplement, which was supposed to handle another difficult part of the *Disquisitiones*, the composition of forms. Again, Gauss

had defined and studied through complicated formulas what it means for three quadratic forms F, f and f' to say that F is composed of f and f'; the concept of composition is inherited by classes (and also genera) of forms and Gauss then uses it critically to prove some important statements of the theory of genera. Dirichlet had simplified the approach by limiting himself to defining a concept of composition for two so-called concordant binary quadratic forms (that is, having the same determinant and coefficients (a, b, c) and (a', b', c') such that a, a', and $b + b'$ are relatively prime). Dedekind presents this approach (Sects. 145–149) and shows that the classes of certain quadratic forms of a given determinant form a group, explicitly linking this terminology to that 'introduced by Galois in algebra' (Sect. 149: the supplement continues the Section numbering of the first edition). He then derives some classical applications to the theory of genera, as well as Gauss's second proof (the third in the *Vorlesungen*) of the law of quadratic reciprocity.

From the following Section on, however, Dedekind's aim is 'to introduce the reader to a higher domain, where algebra and number theory are linked most intimately together [...] The concepts [brought forward here] lead in the algebraic direction towards the principles of Galois, on the arithmetical side towards Kummer's creation of ideal numbers' [Dedekind, 1871, 400–401]. The main concept alluded to here is that of a (number) field (*Körper*), defined in the now familiar way as a system of numbers 'which has the property that the sums, differences, products and quotients of any two of these numbers still belong to the same system' [1871, 400]. The following paragraphs define and study algebraic numbers (which generate these fields), algebraic integers, modules (used to generalize the concept of congruence), and last, but not least, ideals and classes of ideals (Sect. 163). Dirichlet's theorem on the structure of units is set in this perspective (Sect. 166) and a computation is provided for the (finite) number of classes of ideals in a number field (Sect. 167), the supplement ending with some details for the case of quadratic number fields.

According to Dedekind's preface to this 1871 edition, the general theory of ideals was intended 'to throw some light on the main subject of the whole book from a higher perspective'; but he admitted to Rudolf Lipschitz some years later that he 'had believed, that the inclusion of this research in Dirichlet's Zahlentheorie would be the safest means to win a larger circle of mathematicians to work in this field' [Dedekind, *Works*, vol. 3, 464]. His apparent failure, and his own dissatisfaction with the theory, pushed him to rewrite it in the subsequent editions and develop it in a separate autonomous 11th supplement. (On the important technical changes between 1871 and 1894, see Edwards [1980] and Haubrich [1992].) In the last edition, this supplement occupies almost a third of the whole book! Dedekind shared with Dirichlet (and even more explicitly with Riemann) a predilection for conceptual as opposed to computational analyses, but his was increasingly set-theoretical and structural, following the lines of his other foundational writings. Despite Dedekind's early hopes, the complex posterity of his more personal supplements, and their undeniable influence, tended to be almost independent of that of the *Vorlesungen*.

5 THE INFLUENCE OF THE *VORLESUNGEN*

The four successive editions of the *Vorlesungen* alone testify to its success during the last decades of the 19th century. It was not limited to the German-speaking countries (or those,

like Italy, where a translation was made readily available). For instance, Edouard Lucas indicates in his own *Théorie des nombres* of 1891 that the *Vorlesungen* were one of his main sources. In the same vein, George B. Mathews thanked Dedekind in the preface of his introduction on number theory for allowing him 'to make free use of his edition of Dirichlet's *Vorlesungen'*. Browsing through the number-theoretical papers of the 1880s and 1890s reveals recurrent references to some version of the *Vorlesungen* for the basic theorems of the field.

For Dirichlet, 'the variety of methods which serve for the proof of one and the same theorem was a main attraction of number theory' [Dedekind, 1864]; and indeed, the *Vorlesungen* often offered many vistas of the questions treated. Besides the analytical techniques, for instance, one finds in the third supplement an estimate of the connection between the area of a plane figure and the number of lattice points it contains, which would serve as a point of departure for the future geometry of numbers (Sect. 120). Indeed, the variety displayed in the *Vorlesungen* is matched by the variety of the number theorists to whom they appeal. For instance, the copy of the 1871 edition at the *Bibliothèque de mathématiques de Jussieu* in Paris is bound together with reprints and letters of Théophile Pépin and André Desboves, representative authors of the *Nouvelles Annales*, a journal mainly addressed to students, engineers and high-school teachers and defending an elementary approach to number theory.

But, in a striking parallel with Dirichlet's own use of the *Disquisitiones* to educate himself, the *Vorlesungen* will also help to train the younger, more turbulent, generation of number theorists. Around 1880, Heinrich Weber drew the attention of Dedekind to a very promising high school pupil, Minkowski, 'who leaves for the university next year and has worked his way completely on his own in analysis and number theory, which he has studied in the first edition of your Dirichlet–Vorlesungen' [Strobl, 1985, 144]. This involvement would lead Minkowski towards Gauss's *Disquisitiones* and a life-long interest in quadratic forms and the development of the geometry of numbers. In his tribute on Dirichlet for the 100th anniversary of his birth, Minkowski underlines nonetheless, exactly as Hilbert did in the *Zahlbericht*, the contrast between Gauss's and Dirichlet's number theory on one hand and the more recent trends of algebraic number fields, and chooses to present the results of Dirichlet which fit the best into this framework. But in his conclusion, he states [Minkowski, 1905, 162–163]:

> In his lectures Dirichlet treated with predilection these domains, in the construction of which he himself has richly participated. His exposition was thus so penetrating because it appeared as if he was about to create the whole edifice there for the first time; it was captivating to the highest degree to follow him in this work. He developed the material in the most natural way. No artifice occurred as a *deus ex machina* to lead tragically tangled knots to unexpectedly happy solutions. [...] And we today, when we strive more than ever to recognize and represent science in its simple truth, are we not members of the Dirichlet school?

BIBLIOGRAPHY

Dedekind, R. *Works. Gesammelte mathematische Werke*, 3 vols., Braunschweig: Vieweg, 1931–1933. [Repr. in 2 vols. New York: Chelsea, 1969.]

Dedekind, R. 1864. 'Anzeige der ersten Auflage von Dirichlets Vorlesungen über Zahlentheorie', *Göttingische gelehrte Anzeigen*, 121–124. [Repr. in *Works*, vol. 2, 394–395 (cited here).]

Dedekind, R. 1871. 'Anzeige der zweiten Auflage von Dirichlets Vorlesungen über Zahlentheorie', *Göttingische gelehrte Anzeigen*, 1481–1494. [Repr. in *Works*, vol. 2, 399–407 (cited here).]

Edwards, H. 1980. 'The genesis of ideal theory', *Archive for history of exact sciences, 23*, 321–378.

Gauss, C.F. *Works. Werke*, 12 vols., Leipzig: Teubner, 1863-1933. [Repr. Hildesheim: Olms, 1973.]

Gauss, C.F. 1801. *Disquisitiones arithmeticae*, Leipzig: Fleischer. [Repr. as *Works*, vol. 1, 1863. See §22.]

Haubrich, R. 1992. 'Zur Entstehung der algebraischen Zahlentheorie Richard Dedekinds', Dissertation, Georg-Universität Göttingen.

Klein, F. 1926–1927. *Vorlesungen über die Entwicklung der Mathematik im 19. Jahrhundert*, 2 vols., Berlin: Springer. [Repr. in 1 vol. New York: Chelsea, 1967.]

Koch, H. 1998. 'Gustav Peter Lejeune Dirichlet', in *Mathematics in Berlin* (ed. H. Begehr and others), Berlin: Springer, 33–40.

Kummer, E.E. 1860. 'Gedächtnissrede auf Gustav Peter Lejeune Dirichlet', in *Abhandlungen der Königlichen Akademie der Wissenschaften zu Berlin*, 1–36. [Repr. in Dirichlet *Werke*, vol. 2, Berlin: Reimer, 1897 (repr. in 1 vol. New York: Chelsea, 1969), 311–344 (cited here).]

Minkowski, H. 1905. 'Peter Gustav Lejeune Dirichlet und seine Bedeutung für die heutige Mathematik', *Jahresbericht der Deutschen Mathematiker-Vereinigung, 14*, 149–163.

Scharlau, W. (ed.) 1981. *Richard Dedekind 1831–1981. Eine Würdigung zu seinem 150. Geburtstag*, Braunschweig and Wiesbaden: Vieweg.

Smith, H.J.S. 1859–1865. 'Report on the theory of numbers', parts 1–6, in *Report of the British Association for the Advancement of Science* for those years. [Repr. in *The collected mathematical papers*, vol. 1, Oxford: Clarendon Press, 1894 (repr. New York: Chelsea, 1965), 38–364 (cited here).]

Strobl, W. 1985. 'Aus den wissenschaftlichen Anfängen Hermann Minkowskis', *Historia mathematica, 12*, 142–156.

CHAPTER 38

BERNHARD RIEMANN, POSTHUMOUS THESIS ON THE REPRESENTATION OF FUNCTIONS BY TRIGONOMETRIC SERIES (1867)

David Mascré

In this work, prepared for a doctoral defence in 1854 but published only after his death, Riemann both refined the understanding of the integral but especially opened a new era in the handling of Fourier series. His explorations led to new insights into functions and infinite series, and led to the creation of set theory.

First publication. 'Über die Darstellbarkeit einer Funktion durch eine trigonometrische Reihe', *Abhandlungen der Königlichen Gesellschaft der Wissenschaften zu Göttingen, 13* (1867), 87–132. Also Göttingen: Dieterich, 1867.

Manuscript. In Riemann's *Nachlass*, Göttingen University Library Archives.

Reprints. In *Gesammelte mathematische Werke* (ed. H. Weber and R. Dedekind), 1st ed., Leipzig: Teubner, 1876, 213–253. Also in 2nd. ed., 1892, 227–271. [This ed. repr. with additions Berlin: Springer, 1990.]

French translation. In *Bulletin des sciences mathématiques, (1) 5* (1873), 20–48, 79–96. [Repr. in Riemann, *Oeuvres mathématiques* (ed. L. Laugel), Paris: Gauthier–Villars, 1898, 227–279.]

Partial Spanish translation by J. Ferreirós in (ed.), *Riemanniana selecta*, Madrid: Consejo Superior des Investigaciones Cientificas, 2000, 41–60.

Related articles: Cauchy on real-variable analysis (§25), Fourier (§26), Cantor (§46), Lebesgue and Baire (§59).

1 BACKGROUND

Starting with the work of Joseph Fourier (1768–1830) (§26), the question of the representation of functions by trigonometric series constitutes one of the main lines of mathematical

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

analysis through the 19th century. If the works of A.-L. Cauchy, J.P.G. Dirichlet and P. Seidel are milestones in the understanding of the problem, it is in one of *Habilitation* theses (1854) of Bernhard Riemann (1826–1866), presented to the University of Göttingen, that the ideas that contained the bases of the theories can be found. Under an apparently modest title 'On the developability of a function by a trigonometric series', it lays the foundations of what will open a new epoch in real-variable analysis. Three years earlier he had submitted an essay on analytical functions [Riemann, 1851] for his first doctorate (§34) and in 1854 he presented another *Habilitation* thesis 'Über die Hypothesen welche zugrunde der Geometrie liegen' ([Riemann, 1867]: see §39). He did not publish either thesis; they appeared in 1867 under the control of his friend Richard Dedekind (1831–1916).

Composed of three parts, Riemann's thesis on series is a masterpiece of balance and concision. Shuttling between history and reflection, it binds together intrinsic analysis of problems and historical study of their genetic development. It is a perfect example of a thought in which the work of theoretical elaboration is not dissociated from a historic reflection on the origins and genesis of concepts. It is the testimony of a thought where the technical character of demonstrations is masked by the beauty and the depth of intuition.

This genetic and lively conception of mathematical development is perfectly reflected in the choice of the plan of the thesis. The first part (arts. 1–3) treats the history of the representation of an arbitrary trigonometric series by a function. Then follows second part (arts. 4–6) on the study and definition of the notion of integral. The third part, by far the richest, concerns the general study of the representability of function by these series (arts. 7–13).

2 RIEMANN'S HISTORICAL ANALYSIS OF THE INTEGRAL AND TRIGONOMETRIC SERIES

The historical account presented by Riemann is not a simple retrospective. More than a restitution, it is a synthetic overview constituting an authentic conceptual re-appropriation of the ideas discovered by his predecessors. Here history is not separated from thought but intimately linked to it. Its invocation, in the best German university tradition, aims at grasping an intelligent view of the past, in order to find a synthetic and global overview, with the aim of summarizing its logic and to point out its guidelines. His purpose was to enhance the essential novelty of each of the great steps of the development that, notwithstanding the historical incidents, were to lead to the theory at this first stage of achievement where he received it. Riemann articulates these steps around three great historical figures presented as the main actors of this story: Euler, Fourier and Dirichlet. Dirichlet especially had a decisive influence on Riemann, not only by teaching Riemann elements of analysis during his study in Berlin, but as a mentor too, by helping Riemann during the autumn of 1852 to write his thesis on trigonometric series. This is testified by a letter from Riemann to his father: 'The other morning Dirichlet was with me for nearly two hours. He gave me notes that I need for my *Habilitationsschrift* and they are so complete that my work has been made considerably easier. Otherwise I could have spent a long time searching for many things in the library. He also read all my thesis with me and was very kind to me' [Dedekind, 1876, 546].

According to Riemann, Fourier made the most important contribution. His merit is not only due to the priority of his discovery (on this point, he refutes a claim made by S.D. Poisson of priority for J.L. Lagrange) but especially to the depth of his views: 'It was owing to Fourier that the true nature of trigonometric series was recognized in a perfectly correct way; since that time they have been frequently used in mathematical physics for the representation of arbitrary functions, and in each particular case one saw readily that the Fourier series actually converged to the value of the function; but it took a long time for this important theorem to be proved in all generality' (art. 2). Thanks to Fourier, we have access for the first time to an *exact* and *complete* understanding of the *nature* of trigonometric series. It is summarised in Fourier's formulas

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{+\infty} (a_n \cos nx + b_n \sin nx), \quad \text{where } 0 \leqslant x \leqslant 2\pi \tag{1}$$

and

$$\begin{aligned} a_0 &= \frac{1}{\pi} \int_0^{2\pi} f(x), \quad a_n = \frac{1}{\pi} \int_0^{2\pi} f(x) \cos nx \, dx \quad \text{and} \\ b_n &= \frac{1}{\pi} \int_0^{2\pi} f(x) \sin nx \, dx. \end{aligned} \tag{2}$$

The progress did not lay so much in the calculation procedure for the coefficients—already known to Euler and Lagrange—than in the systematic form of the coefficients and the understanding of representability as marked by '='. Indeed, Fourier is the first to consider the whole: f is analysed by means of (2) and synthesised by means of (1). Analysis and synthesis are two indissociable and complementary moments of harmonic analysis.

But when are we in a position to integrate the function and in which cases does the series of coefficients thus obtained converge effectively towards $f(x)$? The question is far from being solved, and a large part of mathematical research revolves surrounds it in the 19th century. If this is right in the specific case dealt with by Fourier, it is a long way from leading to a general and comprehensive theory. The problem lays foremost on the question of knowing under which general conditions the integration formula should be applied: can it be applied to any arbitrary function, or should different classes be distinguished? As Jean Cavaillès remarked [1938, 52]:

> Hence the focus laid, during the whole period, on studies on integration: for Dirichlet, not only the calculation of the coefficients but also their convergence is subordinated to it. The issue is to know which conditions fulfilled by the arbitrary function are sufficient for a certain integration to be possible. What is the result for the behavior of a function of the property to be representable by a trigonometric series? Lastly, is there a one-to-one representation?

Fourier himself did not fully demonstrate the convergence of his series [1822, arts. 415–416], though he showed that for particular functions the series converges towards them. But is it true in general? The urgency of a reply seems all the more important that Fourier himself claimed that 'a trigonometric series, with coefficients thus determined, can

represent any arbitrary function' [Sachse, 1880, 47]. This is a requirement of the theory, linked to the manner itself in which the coefficients were obtained. Cauchy [1827] tried to give a demonstration of this fact. He failed, as he will later admit, but it gives Riemann the opportunity to bring a first precision.

Cauchy had supposed that any periodic function $f(x)$ could be extended into an analytical function $f(x)$ bounded over the whole plane [1827, 603]. As Riemann noticed, this is only true when $f(x)$ is constant. He observes that Cauchy only needed to extend $f(x)$ within the real part of an analytical function $F(x+iy)$, defined and bounded in the upper half-plane $y > 0$, which can be established either by complex methods or by Fourier series. In fact, the proof by complex methods can be found in the thesis ([Riemann, 1851]: see §34). Incidentally, there appears a first connection between complex methods and Fourier series. It is the first step to succession of the coming results, which will ceaselessly confirm the constant interaction between real and complex analysis. This enables Riemann to demonstrate the equivalence between the works started by Fourier and Cauchy (art. 2):

> Cauchy supposes that in a periodic function given arbitrarily, x is replaced by a complex argument $x+iy$, this function is finite for any value of y, but this is only valid if the function $f(x)$ is equal to a constant value. However, it is easily noticed that this supposition is not necessary for further conclusions. It suffices to have a function $\phi(x+iy)$ finite for all positive values of y and whose real part becomes equal, for $y=0$, to the given periodic function $f(x)$. If this supposition, which is indeed right, is first agreed on, the path started by Cauchy leads straight to the aim, as inversely this proposition can be deduced from the theorem on the Fourier series.

But it is about another inaccuracy of Cauchy that Riemann supplies the most interesting contribution. Cauchy's strategy consisted in reducing the study of the convergence of the Fourier series into that of another one, which was easier to study and clearly convergent. According to Cauchy, the Fourier series converges because the ratio between its general term and $(\sin nx)/x$ tends towards 1 when n tends towards ∞. But he had erred by stating that series with terms in such a relation simultaneously converge or diverge. This led Riemann to the fruitful distinction between absolutely convergent series (first class) and conditionally convergent ones (second class). For the latter, showed Riemann, the sum obtained by modifying the order of the terms can equal any finite number. This led him to a fundamental conclusion (art. 3):

> Laws of finite sums may only be applied to series of the first class; those only may be considered as the set of their terms; whereas those of the second class may not. This is a circumstance that had escaped the mathematicians in the last century, mainly for the reason that series that progress according to increasing powers of a variable belong, generally speaking (that is to say to the exception of some specific values of this variable), to the first class.

However, and this is the vital point: 'The Fourier series, obviously, does not necessarily belong to the first class: it was therefore impossible, as Cauchy had vainly tried to do, to deduce its convergence from the law according to which terms decrease'.

This is the explanation of the famous counter-example exhibited in [Dirichlet, 1829] against Cauchy's argument. Dirichlet had rightly noted 'that it is easy to give the example of series whose terms are not all positive, such that one is convergent, the other divergent, but that the ratio of the corresponding terms tends to $+1$' [1829, 158], and offered the example

$$\sum_{n=0}^{+\infty} \frac{(-1)^n}{\sqrt{n}} \quad \text{and} \quad \sum_{n=0}^{+\infty} \frac{(-1)^n}{\sqrt{n}} \left[1 + \frac{(-1)^n}{\sqrt{n}} \right]. \tag{3}$$

The correct method, as Dirichlet had already showed, lay in studying the convergence of the sum of the first n terms in the Fourier series, or, and this amounts to the same, the integral

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} f(\alpha) \frac{\sin \frac{2n+1}{2}(x-\alpha)}{\sin \frac{x-\alpha}{2}} \, d\alpha \tag{4}$$

when n tends towards infinity. In the case where the function $f(\alpha)$ is continuous, monotonic and finite, the integral obviously converges towards $f(x)$. But Dirichlet had gone even further by succeeding in demonstrating, on the basis of the rules already identified by Cauchy, that the result could be extended to functions monotonic by parts. To do this, all he had to do was to invoke the additivity of the integral and the rule of convergence under the summation sign:

1) The integral of a function within an interval being the sum of the integrals over a finite number of partial intervals in which the first one is subdivided, f can take a finite number of minima and maxima.

2) If the integral $\int_a^x f(t)\,dt$ converges when x tends towards b, a singular point for f (i.e. point where f is either discontinuous or infinite), then the integral tends towards a determined limit, equal to $\int_a^b f(t)\,dt$. In particular, in the case when f has a right-hand and a left-hand limit in b, the two integrals $\int_a^{b-0} f(t)\,dt$ and $\int_{b+0}^c f(t)\,dt$ are correctly defined and

$$\int_a^c f(t)\,dt = \frac{1}{2}\left\{ \int_a^{b-0} f(t)\,dt + \int_{b+0}^c f(t)\,dt \right\}. \tag{5}$$

With those results in mind, Dirichlet's theorem became immediately obvious. One can represent by a trigonometric series each 2π-periodic function that remains periodic in the interval $[-\pi, \pi]$ and does not have infinitely many discontinuities or maxima and minima.

Riemann summarised Dirichlet's proof by reasoning directly on his integral. The proof was based on the following facts:

$$\lim_{n\to\infty} \frac{1}{2\pi} \int_0^c f(\beta) \frac{\sin(2n+1)\beta}{\sin\beta} \, d\beta = \frac{\pi}{2} f(0) \quad \forall c \text{ such that } 0 < c \leqslant \frac{\pi}{2}; \tag{6}$$

and

$$\lim_{n\to\infty} \frac{1}{2\pi} \int_b^c f(\beta) \frac{\sin(2n+1)\beta}{\sin\beta} \, d\beta = 0 \quad \forall b, c \text{ such that } 0 < b < c \leqslant \frac{\pi}{2}, \tag{7}$$

where the function f is supposed to be always increasing or always decreasing within the limits of these intervals (art. 3):

Indeed, these two conditions suffice, in the case where the function does not change an infinite number of times from increasing to decreasing steps, to decompose the integral

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} f(\alpha)\frac{\sin\frac{2n+1}{2}(x-\alpha)}{\sin\frac{x-\alpha}{2}}\,d\alpha$$

into a finite number of terms, one of which converges towards $\frac{1}{2}f(x+0)$, another towards $\frac{1}{2}f(x-0)$, and all the others towards 0, when n increases to the infinite.

This provides an explanation of Dirichlet's result, which Riemann summarized as follows (art. 3):

A trigonometric series can represent every periodic function with the period 2π which

1) is integrable throughout,
2) does not have infinitely many maxima and minima,
3) takes on the mean value of its two limiting values wherever its value changes abruptly, that is, it is such that $f(x) = (f(x+0) + f(x-0))/2$.

This last condition, added by Riemann, is in fact necessary. Indeed, 'a function which has the first two properties and not the third can obviously not be represented by a trigonometric series: the trigonometric series which would represent it outside the discontinuities would differ from it at the point of discontinuity itself' (art. 3). For if condition 2) is not valid, the conclusion about Dirichlet's integral is not valid; if instead 1) is dropped, where the integral is understood in Cauchy's sense, it is not possible to determine the coefficients of the Fourier series. However, none of the first two given conditions is irreducible, as Dirichlet had already noted [1829, 169]:

> we would still have to consider the case where the suppositions we have made on the number of solutions of continuity and on the number of minima and maxima values cease to occur. These particular cases can be compared to those we have just considered [...] But the point, if it is to be performed with all the desired clarity, requires some details linked to the fundamental principles of infinitesimal analysis, which will be exposed in another note.

This finding implied a general revision of the theory of integration, a promise which neither Dirichlet nor his successors made. Hence a mixed feeling, with a hint of disappointment, clouds Riemann's final judgement: 'Knowing if and when a function which does not fulfil the first two conditions is representable through a trigonometric series, this is what remains open in Dirichlet's research' (art. 3).

Nevertheless, the theory carried the first marks of achievement—which was all the more considerable in that, for Riemann, it was destined to cover most of nature's phenomena (art. 3):

> This work of Dirichlet has given a solid basis to a great number of important analytical researches. By highlighting a point on which Euler had erred,

> he has succeeded in solving a question which had bothered so many eminent mathematicians for over seventy years (since 1753). In fact, the problem was completely solved for all cases which present themselves in nature alone, because however great may be our ignorance about how the forces and states of matter vary in space and time in the infinitely small, we can certainly assume that functions to which Dirichlet's research did not extend do not occur in nature.

We might be surprised by the manner in which Riemann puts aside the possibility of such a case just after he has considered extending his results to functions more irregular than those specified by Dirichlet. But Riemann still points out other possible extensions, as fundamental and even more unexpected: those which enable to connect the theory of Fourier series not only to physics but also to the theory of numbers.

The connection is all the more important that it shows the existence of connections as profound as unexpected between analysis and arithmetic [Knobloch, 1983, 323]. This was a fine idea, of which G.W. Leibniz, Leonhard Euler and Carl Jacobi had already given some illustrations, with numerous repercussions; for example, for Riemann a few years later when he discovered one of the keys to the theory of theta functions.

3 RIEMANN ON THE INTEGRAL

Too closely dependent on the definition of the integral given by Cauchy, the path opened up by Dirichlet could therefore not be pursued further, without triggering a profound revision of the concept. Here was the evidence of an internal obstruction, directly linked to the uncertainties that were still reigning over certain fundamental points of the infinitesimal calculus. This led Riemann to his crucial question: 'what do we understand by $\int_a^b f(x)\,dx$?' (art. 4).

The response that Riemann gives to that question consists of the definition of the integral that carries his name. This is the topic of the second part of the thesis. 'The positive part of Riemann's thesis starts indeed with a new definition of the integral, whose properties enable him to show both one of the fundamental lemmas, and directly, some of the main results, such as the theorem stating that when the function is integrable in this new sense, the coefficients of the new trigonometric series effectively converge towards zero' [Cavaillès, 1962, 53]. The originality of Riemann's approach lays in the transformation which it operates, as rightly pointed out by Lebesgue [1906], by turning the operating process given by Cauchy to calculate the integral of a continuous function into the starting point of the definition of the integral of any function [Hawkins, 1970, ch. 1]. 'Cauchy's definition (which was also adopted in his work of 1829) was applicable when the integrand $f(x)$ is continuous in the interval of integration or presents at most a finite number of points of discontinuity. However it ceased to be valid if these points are infinite in number. But it was precisely this case that interested Riemann' [Sachse, 1880, 243]. By treating it, Riemann gave a first illustration of the principle of conservative extensions, which in fact enabled him to clearly define the notion of integral.

Riemann's method of defining the integral of a function between two points a and b was to divide the interval $[a, b]$ into n parts $[x_{i-1}, x_i]$ with $a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b$; to consider in each of the partitions thus formed any point $y_i = x_{i-1} + \varepsilon_i \delta_i$ (where $0 < \varepsilon_i < 1$ and $\delta_i = x_i - x_{i-1}$ is the difference between two successive values); and to examine how the sum $S = \sum_{i=1}^{n} \delta_i f(y_i)$ thus obtained depends on the choice of the partial intervals δ_i and fractions ε_i. Then 'the value of this sum will depend on the choice of the intervals δ and of the magnitudes ε. If it has the property that, regardless of the choice of δ and ε, it approaches indefinitely a fixed limit A as all δ's become infinitely small, then this value is called $\int_a^b f(x)\,dx$'. But if the sum S tends towards no limit, the notation $\int_a^b f(x)\,dx$ 'cannot have any meaning'. A natural extension of this definition occurs when, as $f(x)$ becomes infinitely large as x tends to a value c, there nevertheless exists the limit of $\int_a^{c-\alpha_1} f(x)\,dx + \int_{c+\alpha_2}^b f(x)\,dx$ as α_1 and α_2 tend to zero (art. 4).

This new point of view allowed an effective extension of the notion of integral of Cauchy, in the sense that some functions to which Cauchy's definition is not applicable are Riemann integrable. It also led to a determination of the necessary and sufficient conditions for the existence of the integral for a finite-valued function f. Considering the oscillation ω_i of the function within interval δ_i (that is, the difference between the upper and lower limit of the values taken by this function over this interval), Riemann noticed that the necessary and sufficient condition for S to have a limit, that is for f to be integrable, is that 'the total sum of intervals δ_i in which the oscillations ω_i are $> \sigma$, regardless of the value of σ, can be made as arbitrarily small by an appropriate choice of d $[= \sup \delta_i]$' (art. 5). Fifty years later in his *Leçons sur l'intégration*, Henri Lebesgue will discover the necessary and sufficient condition for the integrability of functions, and demonstrate that a function is integrable in the Riemann sense if and only if its set of points of discontinuity is of measure zero (§59.6).

Meanwhile, Riemann's definition supplied an adequate instrument for a kind of generalisation which neither Dirichlet nor (after Riemann) [Lipschitz, 1864] had managed to reach. In fact, even a series possessing a dense set of discontinuities can be integrable. Riemann gives the example of the series

$$f(x) = \sum_{n=1}^{+\infty} \frac{(nx)}{n^2} \quad \text{where } (x) = \begin{cases} x - m & \text{if } |x - m| < 1/2 \\ 0 & \text{when } x = m + 1/2; \end{cases} \tag{8}$$

it is convergent at every point and continuous everywhere except at points $x = p/(2n)$ (an irreducible fraction) where it leaps by $\pi^2/(16n^2)$ (art. 6). The discontinuities form a dense set, but there are only finitely many jumps $> h$ in every finite interval, as can easily be seen by noting that

$$\left\{ n\colon \frac{\pi^2}{16n^2} > h \right\} = \left\{ n\colon n < \sqrt{\frac{\pi^2}{16h}} = \frac{\pi}{4} h^{-1/2} \right\}. \tag{9}$$

So the function is integrable.

4 THE PROBLEM OF THE UNIQUENESS OF THE REPRESENTATION

Having refined the concept of integrability, Riemann took up to the central point of his research: the determination of the representability of functions by Fourier series. Preceding studies, he says, have followed this scheme: if a function has such and such properties, then it can be expanded in Fourier series. But 'We must proceed from the inverse question: if a function is representable by a trigonometric series, what consequences does this have for its behavior, for the variation of its value with the continuous variation of the argument?' (art. 7). He was the first to distinguish clearly the problem of representing a given function $f(x)$ by a trigonometric series from its converse problem: that is, explaining the consequences for the behavior of $f(x)$ as x changes continuously of its representability by a trigonometric series. Dirichlet's finding gave Riemann a sufficient condition for the first problem; Riemann discovered necessary conditions for the second. The question, today still open, of finding necessary and sufficient conditions for a 2π-periodic function $f(x)$ to be equal for all real x to its Fourier series, was probably one of his major aims.

In order to do so, Riemann considered the following series

$$\Omega = \sum_{n=0}^{+\infty} A_n = \frac{a_0}{2} + \sum_{n=1}^{+\infty} (a_n \sin nx + b_n \cos nx). \tag{10}$$

For each value of x for which the series Ω converges, he sees that $f(x)$ is its limit. Recalling a characterisation given by Cauchy [1827], he notes that a necessary condition for Ω to converge is that $\lim_{n\to\infty} A_n = 0$. Two cases may then occur, depending on the fact that $\lim_{n\to\infty} A_n = 0$ for all values of x or at the exception of some of them.

In the first case, the Fourier coefficients converge to zero. (Formally, this means that

$$\lim_{n\to\infty} (a_n \sin nx + b_n \cos nx) = 0 \quad \forall x \in [a,b] \quad \Rightarrow \quad \lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n = 0, \tag{11}$$

which will only be proved by Cantor: see [Cooke, 1993], and compare §46.) The key to the demonstration lies then in the exhibition of an auxiliary function. Riemann takes it as

$$F(x) = C + C'x + \frac{A_0 x^2}{2} + \sum_{n=1}^{+\infty} \frac{A_n}{n^2}, \tag{12}$$

which is obtained by formally integrating twice Ω. The advantage of this function is that it allows a simultaneous treatment of the problem of the representation and the problem of the convergence of the series. The function is clearly continuous and integrable. Three lemmas then enable him to establish the necessary condition for the representation (art. 8):

LEMMA 1. *If the series converge to $f(x)$ and if α and β decrease to zero in such a way that their ratio remains finite, then*

$$\frac{F(x+\alpha+\beta) - F(x+\alpha-\beta) - F(x-\alpha+\beta) + F(x-\alpha-\beta)}{4\alpha\beta} \tag{13}$$

converges towards $f(x)$.

LEMMA 2. *The function*

$$\frac{F(x+2\alpha)+F(x-2\alpha)-2F(x)}{2\alpha} \tag{14}$$

converges towards zero as α tends towards zero.

LEMMA 3. *Designate by b and c two arbitrary constants such that $b<c$ and by $\lambda(x)$ a function continuous between b and c, whose first derivative has the same properties and whose second derivative does not have an infinite number of maxima and minima. Then*

$$\lim_{n\to\infty} n^2 \int_b^c F(x)\cos n(x-a)\lambda(x)\,dx = 0. \tag{15}$$

This theorem has become known as his 'localisation theorem'.

Riemann now focusses upon representability (art. 9). He obtains two theorems:

THEOREM 1. *If a 2π-periodic function $f(x)$ can be represented by a trigonometric series whose terms ultimately became infinitely small for every value of x, then there must be a continuous function $F(x)$, on which $f(x)$ depends, such that*

$$\frac{F(x+\alpha+\beta)-F(x+\alpha-\beta)-F(x-\alpha+\beta)+F(x-\alpha-\beta)}{4\alpha\beta} \tag{16}$$

converges towards $f(x)$ when α and β become infinitely small and their ratio remains finite. Moreover,

$$\lim_{n\to\infty} n^2 \int_b^c F(x)\cos n(x-a)\lambda(x)\,dx = 0, \tag{17}$$

where $\lambda(x)$ is assumed to be a function continuous between b and c and null at b and c, whose first derivative has the same properties, and whose second derivative does not have an infinite number of maxima and minima.

THEOREM 2. *Conversely, if these two requirements are fulfilled, then there exists a trigonometric series in which the coefficients ultimately and which represents the function wherever the series converge.*

THEOREM 3. *Let $b<x<c$ and let $r(t)$ be a function such that $r(t)=r'(t)=0$ for $t=b$ and $t=c$, and such that $r(t)$ and $r'(t)$ varies continuously between these values. Suppose moreover that $r''(t)$ does not have infinitely many maxima and minima and that $r(t)=1$, $r'(t)=0$, $r''(t)=0$ for $t=x$ while $r'''(t)$ and $r''''(t)$ are finite and continuous. Then the difference between the series*

$$\Omega = \sum_{n=0}^{+\infty} A_n = \frac{b_0}{2} + \sum_{n=1}^{+\infty}(a_n \sin nx + b_n \cos nx) \tag{18}$$

and the integral

$$\frac{1}{2\pi}\int_b^c F(t)\frac{d^2(\sin\frac{2n+1}{2}(x-t))/\sin\frac{x-t}{2}}{dt^2}r(t)\,dt \tag{19}$$

converges towards zero when n increase indefinitely. Therefore, the series (18) *converges or not according to whether Ω approaches a fixed limit or not as n increases indefinitely.*

In this way, studying the general convergence of the series reduces to examining the behaviour of a particular integral. Theorems 1 and 2 give the necessary conditions for a given function $f(x)$ to be represented by a trigonometric series that has $f(x)$ as its generalized Riemann sum. In fact, $f(x)$ is the generalized second derivative of $F(x)$ as given in (16). The result is the more impressive because Riemann did not frame any hypothesis about the form of the coefficients in the series (18). Consequently, the results remain true even when the coefficients are not coefficients of Fourier series. The consequences of the distinction between Fourier series and trigonometric series will first appear clearly with [Cantor, 1872]; see also [du Bois-Reymond, 1880].

In art. 12 Riemann considers the second case, where the coefficients of Fourier become infinitely small with $1/n$ for a value of x without this happening for all values. In that case the series does not necessarily converge towards zero for all x. But the result remains valid. To see this, it suffices to substitute $x+t$ and $x-t$ in the definition of Ω. Summing term by term, we indeed obtain the series whose terms tend to zero as n increases for every value of t and to which we can therefore apply the results obtained.

5 THE FINAL ARTICLE: EXAMPLES ILLUSTRATING THE DIVERSITY AND COMPLEXITY OF TRIGONOMETRIC SERIES

Riemann did not consider all cases that are excluded from Dirichlet's conditions, but limited himself to a few important examples. Art. 13 is in that sense a kind of final fireworks, where Riemann raises more problems than he was apparently able to solve. As Laugwitz put it, 'Special examples serve here to investigate the scope of concepts', especially of continuity and piecewise differentiability [Laugwitz, 2000, 188].

As far as functions with an infinite number of maxima and minima are concerned, Riemann asserts that there are functions that are integrable in the wide sense that are not representable by Fourier series: for instance,

$$f(x)=\frac{d}{dx}\left(x^{\nu}\cos\frac{1}{x}\right),\quad \text{where } 0<\nu<\frac{1}{2}, \tag{20}$$

which satisfies

$$\int_0^{2\pi} f(x)\cos n(x-a)\,dx\approx\frac{1}{2}\sin\left(2\sqrt{n}-na+\frac{\pi}{4}\right)\sqrt{\pi}n^{(1-2\nu)/4} \tag{21}$$

and whose series of Fourier coefficients is therefore divergent. This proves that a Fourier series of an integrable function in the wide sense may be divergent. Inversely, following a

comment probably considered by Riemann, but which Fatou [1906] will later demonstrate, a trigonometric series may be everywhere convergent without its sum being integrable on any interval. For instance, the function

$$f(x) = \sum_{n=2}^{+\infty} \frac{\sin nx}{\log n} \tag{22}$$

is everywhere convergent, whereas its sum is neither Riemann nor Lebesgue integrable [Lebesgue, 1906, 124].

Riemann then considers non-integrable functions, whose associated series may converge on a dense set. The result is a multitude of equally fundamental examples. The first one is the function

$$f(x) = \sum_{n=1}^{+\infty} \frac{(nx)}{n} \tag{23}$$

(where the function (x) represents the difference between x and the closest integer), defined for every rational value of x and which is representable (which Riemann just indicates without proving it) by

$$f(x) = \sum_{n=1}^{+\infty} \frac{d'_n - d''_n}{n\pi} \sin 2\pi nx, \tag{24}$$

where d'_n is the number of odd divisors and d''_n the number of even divisors of n. This function exists almost everywhere and its oscillation is infinite over any interval, so it is nowhere Riemann integrable. It is nevertheless Lebesgue integrable, and the previous expression is none other than its development as a Fourier–Lebesgue series. It links the theory of series to problems of arithmetic, and may have suggested to Cantor of the importance of an arithmetical approach to analysis (see Cantor [1883, arts. 4 and 10]; and compare §46).

Riemann then considers even stranger examples such as the series $\sum_{n=1}^{+\infty} \frac{(nx)}{n^2}$. All of its jumps are negative and the sum is infinite, so that it is not Riemann integrable; but the continuous parts of the function (nx) gives a positive derivative which ensures its convergence.

Probably independently of Riemann, [Jordan, 1881] will deduce the existence of functions with bounded variation and however discontinous on any interval, for instance the series $\sum_{n=1}^{+\infty} \frac{(nx)}{n^3}$. Similarly, Riemann gives two series, $\sum_{n=1}^{+\infty} c_n \cos n^2 x$ and $\sum_{n=1}^{+\infty} c_n \sin n^2 x$, where the c_n are positive quantities decreasing to zero, but for which the series $\sum_{n=1}^{+\infty} c_n$ diverges.

Lastly, Riemann takes $\sum_{n=1}^{+\infty} \sin n!\pi x$, which converges at all rational points (as well as in some irrational points, which he tries to show) whereas its coefficients do not tend towards zero. This shows that the trigonometric series can also converge infinitely between any two arbitrarily close arguments if its coefficients do not ultimately become infinitely small.

At the same time, Riemann wonders about the existence of continuous functions that may be nowhere differentiable. As testified by [Weierstrass, 1872], he gave $\sum_{n=1}^{+\infty} \frac{\sin n^2 x}{n^2}$

to his students as a possible candidate. It was indeed logical, when considering the formal derivative of this function $\sum_{n=1}^{+\infty} \cos n^2 x$, to see there the limit of the Gauss sums $\sum_{n=1}^{+\infty} \exp[in^2(x+iy)]$ $(y>0)$. But neither Riemann nor Weierstrass was able to establish rigorously the expected result. In fact, the proof will only be given in [Hardy, 1916], where the non-differentiability of the function is established, first over the irrationals, then over a certain class of rationals, then lastly at any point other than those of the form $x = \pi \frac{2p+1}{2q+1}$, p and q integers [Gerver, 1971, 32–55].

Then Weierstrass [1872] shows another example of such a function,

$$F(x) = \sum_{n=1}^{+\infty} b^n \cos a^n \pi x, \tag{25}$$

where a is a sufficiently large odd integer. For b positive and strictly smaller than 1 $(0 < b < 1)$, the series is uniformly convergent, and the function thus defined is therefore continuous. If we have $ab < 1$, the derivative of $F(x)$ is

$$F'(x) = -\pi \sum_{n=1}^{+\infty} (ab)^n \sin a^n \pi x; \tag{26}$$

but if $ab > 1 + 3\pi/2$, $F(x)$ no longer has a derivative. Indeed, the existence of a derivative would require that the differential quotient $\Delta = |(F(x+h) - F(x))/h|$ remains inferior to ε for all values $|h| < \alpha$, ε being arbitrary and α a given number. Fairly simple transformations show that, in this given case, the hypothesis

$$|h| < \frac{3}{2a^m} \quad \text{leads to} \quad \Delta > \frac{3}{2a^m}\left(\frac{2}{3} - \frac{\pi}{ab-1}\right)(ab)^m. \tag{27}$$

Thus $\lim_{h\to\infty} \Delta = +\infty$, and the function $F(x)$ is nowhere differentiable. In fact the conditions $a < 1$ and $ab \geqslant 1$ suffice to ensure non-differentiability [Hardy, 1916].

We stress the beauty of this final part: real fireworks of examples, a goldmine of material where later mathematicians will continuously dig: in a way, most of real-variable analysis of the later 19th and 20th centuries has sprung from here in one way or another. But Riemann was not in a position to deal in a general manner with functions containing an infinite number of maxima and minima; for their definition would have required that, for a given value σ, a partitioning d could always be chosen in order to ensure that the conditions for the integrability were effectively satisfied. On this issue, his thesis left a number of points unresolved, which is presumably why he decided to postpone its release (the rather scrappy form of art. 13 suggests this hypothesis). It remains that its publication in 1867 gradually opened the way to an incredible number of new theories: definition of the integral, characterization of integrable functions and introduction of sets of zero Lebesgue measure, study of general trigonometric series, formal integration, relations between real and complex methods in Fourier analysis, fractal objects, scale factors, pseudo-functions and smooth functions, oscillating integrals, and condensation of singularities: these are all concepts and techniques that directly issued from Riemann's idea and founded what will become in the next decades the vast field of modern real-variable analysis.

BIBLIOGRAPHY

Bottazzini, U. 1986 *The higher calculus. A history of real and complex analysis from Euler to Weierstrass*, New-York: Springer.

Burkhardt, H. 1908. 'Entwicklungen nach oscillirenden Functionen und Integration der Differentialgleichungen der mathematischen Physik', *Jahresbericht der Deutschen Mathematiker-Vereinigung, 10*, 1804–5.

Cantor, G. *Papers. Gesammelte Abhandlungen* (ed. E. Zermelo), Berlin: Springer, 1932.

Cantor, G. 1872. 'Über die Ausdehnung eines Satzes aus der Theorie der trigonometrischen Reihen', *Mathematische Annalen, 5*, 23–32. [Repr. in *Papers*, 92–102.]

Cantor, G. 1883. *Grundlagen einer allgemeinen Mannigfaltigkeitslehre*, Leipzig: Teubner. [Repr. in *Papers*, 165–208. See §46.]

Cauchy, A.L. 1827. 'Mémoire sur les développements des fonctions en séries périodiques', *Mémoires de l'Académie Royale des Sciences, 6* (1823), 603–612. [Repr. in *Œuvres complètès*, ser. 2 vol. 1, Paris: Gauthier-Villars, 1908, 12–19.]

Cavaillès, J. 1938. *Remarques sur la formation de la théorie des ensembles*, Paris: Hermann.

Cavaillès, J. 1962. *Philosophie mathématique*, Paris: Hermann.

Cooke, R.L. 1993. 'Uniqueness of trigonometric series and descriptive set theory', *Archive for history of exact sciences, 45*, 281–334.

Darboux, G. 1875. 'Mémoire sur les fonctions discontinues', *Annales scientifiques de l'Ecole Normale Supérieure, (2) 4*, 57–112.

Dauben, J.W. 1971. The trigonometric background to Georg Cantor's theory of sets', *Archive for history of exact sciences, 7*, 181–216.

Dauben, J.W. 1979. *Georg Cantor. His mathematics and philosophy of the infinite*, Cambridge (Mass.) and London: Harvard University Press.

Dedekind, R. 1876. 'Riemanns Lebenslauf', in Riemann *Works*, 1st ed., 507–526. [Repr. in 2nd ed. (1892), 539–558 (cited here).]

Dirichlet, J.P.G. Lejeune-. 1829. 'Sur la convergence des séries trigonométriques qui servent à représenter une fonction arbitraire entre les limites données', *Journal für die reine und angewandte Mathematik, 4*, 157–169. [Repr. in *Werke*, vol. 1, Berlin: Reimer, 1889, 283–306.]

du Bois-Reymond, P. 1880. *Zur Geschichte der trigonometrischen Reihen, eine Entgegnung*, Tübingen: Laupp.

Fatou, P. 1906. 'Séries trigonométriques et séries de Taylor', *Acta mathematica, 30*, 335–400.

Fourier, J. 1822. *Théorie analytique de la chaleur*, Paris: Firmin Didot. [Repr. as *Œuvres complètes*, vol. 1, Paris: Gauthier–Villars, 1888. See §26.]

Gerver, J. 1971 'The differentiability of the Riemann function at certain rational multiples of π', *American journal of mathematics, 92*, 35–55.

Grattan-Guinness, I. 1970. *The development of the foundation of mathematical analysis from Euler to Riemann*, Cambridge, MA: MIT Press.

Grattan-Guinness, I. 1980. (Ed.), *From the calculus to set theory 1630–1910*, London: Duckworth.

Hardy, G.H. 1916. 'Weierstrass's non-differentiable function', *Transactions of the American Mathematical Society, 17*, 301–325. [Repr. in *Collected papers*, vol. 4, Oxford: Clarendon Press, 1968, 477–501.]

Hawkins, T.W. 1970. *Lebesgue's theory of integration. Its origins and development*, Madison and London: University of Wisconsin Press.

Jordan, C. 1881. 'Sur la série de Fourier', *Comptes rendus de l'Académie Royale des Sciences, 92*, 228–230. [Repr. in *Œuvres*, vol. 4, Paris: Gauthier–Villars, 1964, 393–395.]

Knobloch, E. 1983. 'Von Riemann zu Lebesgue. Zur Entwicklung der Integrationstheorie', *Historia mathematica, 10*, 318–343.

Laugwitz, D. 2000. *Bernhard Riemann*, Berlin: Birkhäuser.

Lebesgue, H. 1906. *Leçons sur les séries trigonométriques*, Paris, Gauthier–Villars. [See §59.]

Lipschitz, R. 1864. 'De explicatione per series trigonometricas instituenda functionum unius variabilis arbitrarium, et praecipue earum, quae per variabilis spatium finitum valorum maximorum et minimorum numerum habent infinitum, disquisitio', *Journal für die reine und angewandte Mathematik, 63*, 296–308. [French trans.: *Acta mathematica, 36* (1913), 281–295.]

Riemann, G.F.B. *Works. Gesammelte mathematische Werke* (ed. H. Weber and R. Dedekind) 1st ed., Leipzig: Teubner, 1876. [2nd ed. 1892; 3rd ed. with additions Berlin: Springer, 1990.]

Riemann, G.F.B. 1851. 'Grundlagen der allgemeine Theorie der Functionen einer veränderlichen complexen Grösse. (Inaugural-dissertation)', Göttingen. [Repr. in *Works*, 2nd ed., 3–48. See §34.]

Riemann, B. 1867. 'Ueber die Hypothesen, welche der Geometrie zugrunde liegen. (Habilitationsvortrag)', *Abhandlungen der Königlichen Gesellschaft der Wissenschaften zu Göttingen, 13*, 133–152. [Repr. in *Works*, 2nd ed., 272–287. See §39.]

Sachse, A. 1880. 'Versuch einer Geschichte der Darstellung willkürlicher Funktionen einer Variabeln durch trigonometrische Reihen', *Abhandlungen zur Geschichte der Mathematik, 3*, 229–276. [French trans. in *Bulletin des sciences mathématiques, (2) 4*, pt. 1 (1880), 43–64 and 83–112.]

Weierstrass, K. 1872. 'Über continuierliche Funktionen eines reellen Arguments . . .', manuscript, in *Mathematische Werke*, vol. 2, Berlin: Reimer, 1895 (repr. Hildesheim: Johnson; New York: Olms, 1967), 71–74.

CHAPTER 39

BERNHARD RIEMANN, POSTHUMOUS THESIS 'ON THE HYPOTHESES WHICH LIE AT THE FOUNDATION OF GEOMETRY' (1867)

Jeremy Gray

Riemann's lecture, given in 1854 and published posthumously in 1867, is one of the key work from which derives the modern study of differential geometry, and especially the study of manifolds of dimension greater than two. It was to prove central to the overthrow of Euclidean geometry as the source of geometrical ideas and to Einstein's general theory of relativity after 1915.

First publication. 'Ueber die Hypothesen, welche der Geometrie zu Grunde liegen', *Abhandlungen der Königlichen Geselleschaft der Wissenschaften zu Göttingen, 13* (1867), *mathematische Klasse*, 133–152. Also Göttingen: Dieterich, 1867.

Manuscript. In Riemann's *Nachlass,* Göttingen University Library Archives.

Reprints. In *Gesammelte mathematische Werke* (ed. R. Dedekind and H. Weber), Leipzig: Teubner, 1876, 254–269. [Repr. in 2nd ed. (ed. H. Weber), 1892, 272–287. Also in 3rd ed. (ed. R. Narasimhan), Berlin: Springer, 1990, 304–319 (cited below).]

Further reprint. In Riemann, *Ueber die Hypothesen, welche der Geometrie zu Grunde liegen* (ed. and introd. H. Weyl), Berlin: Springer, 1919. [Repr. 1923.]

French translation by J. Houël in *Annali di matematica, (2) 3* (1870), 309–327. [Repr. Paris: Hermann, 1898; also in *Œuvres mathématiques de Riemann* (trans. L. Laugel), Paris: Gauthiers–Villars, 1898 (repr. Paris: Blanchard, 1968), 280–299.]

English translations. 1) By W.K. Clifford in *Nature, 8* (1873), 114–117, 136–137. [Repr. in Clifford, *Mathematical papers* (ed. R. Tucker), London: Macmillan, 1882 (repr. New York: Chelsea, 1968), 55–72. Also in W. Ewald (ed.), *From Kant to Hilbert,* 2 vols., Oxford: Oxford University Press, 652–661.] 2) Trans. G.B. Halsted, in *Tokyo sagaku*

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

bitsurigaku kwai kiji, 8 (1895), 65–78. 3) By and in M. Spivak, *A comprehensive introduction to differential geometry*, vol. 2, Boston, MA: Publish or Perish, 1976, vol. 2, 135–153.

Polish translation by S. Dickstein in *Commentarii Academiae Litterariae Cracov, 9* (1877).

Russian translation by D. Sintsov in *Memoirs of the Physical–Mathematical Society of the University of Kazan, (2) 3* (1893), appendix.

Spanish translation by J. Ferreirós in (ed.), *Riemanniana selecta*, Madrid: Consejo Superior des Investigaciones Cientificas, 2000, 2–18.

Related articles: von Staudt (§33), Poncelet (§27), Klein (§42), Einstein (§63).

1 BERNHARD RIEMANN (1826–1866)

Bernhard Riemann was the son of a German pastor. He led a sheltered life, and originally intended to follow his father into the Church; but his extraordinary talent for mathematics soon took him in that direction, and he studied mathematics in Göttingen and Berlin from 1846 to 1851. His doctoral thesis, published in 1851, is one of the founding texts in the field of complex analysis (§34), and in a series of papers that followed he established himself as one of the most profound conceptual thinkers in mathematics in the entire 19th century. He became an associate professor at Göttingen in 1857 and a full professor in 1859. His influence continues to be felt today, not least in the subject of Riemannian differential geometry, which derives from the thesis under discussion here. Indeed, it could be argued that Riemann's lecture 'On the hypotheses that lie at the basis of geometry' did more to change our ideas about geometry and physical space than any work on the subject since Euclid's *Elements*. Yet by the time it was published in 1867, as one of three of Riemann's papers in the *Göttingen Nachrichten*, its author had died of pleurisy in Selasca, near Lake Maggiore in Italy, in 1866, at the age of 39.

2 THE LECTURE

This posthumous work was originally given as a lecture to the Philosophy Faculty of the University of Göttingen in 1854, in partial fulfilment of the requirements for the award of a *Habilitation*, the German qualification needed before one could teach at a German university. Candidates had also to submit a written thesis, and to offer three topics for a lecture. Riemann offered this title as the third of his list of three, and did not expect to be called to speak on it; but the senior examiner was Carl Friedrich Gauss (1777–1855), and geometry had been a life-long interest of his. Gauss was to announce himself very pleased with what he heard.

The essay opens with a remark about a darkness that lies at the foundations of geometry, and which, in Riemann's opinion, is not illuminated by the usual axiomatic presentation. This darkness obscures the connections between what is assumed, which is the notion of space and of constructions in space. It has persisted from Euclid to A.M. Legendre, to name only the most famous of recent authorities, perhaps because the idea of multiply extended

magnitudes has not been discussed. Once this is done, said Riemann, it will be seen that even among three-dimensional extended magnitudes there is no unique choice, and so the nature of space itself becomes empirical.

This is a remarkably bold opening, promising nothing less than the overthrow of Euclidean geometry as the source of all geometric ideas. It is a challenge not just to the scope and reach of mathematics in 1854, but to whatever philosophy orthodoxy might have prevailed in Göttingen at the time; as we shall see, Riemann had a deep immersion in German philosophical thought of the day.

A multiply-extended magnitude—or manifold ('*Mannigfaltigkeit*'), as he also called it—is not an unintuitive concept. A one-fold extended magnitude, or one-dimensional manifold, is a curve. Points on a curve require one measurement or coordinate to determine their position. If a curve moves along a line it sweeps out a surface or two-dimensional manifold, in which points require two coordinates to specify their position. Riemann said that examples of multiply-extended magnitudes are the positions of perceived objects and colours. There is no need to stop with three-dimensional manifolds, and indeed Riemann contemplated extended magnitudes of arbitrary multiplicity. So, roughly speaking and without looking for complications, a multiply-extended magnitude is something that is captured or measured by using a certain number of coordinates.

Once a position of a point in an n-dimensional manifold is given by stating its n coordinates, the question arises of determining its distance from any other point in the manifold. Evidently this must be measured along a path that lies entirely in the manifold. Riemann noted that Gauss had shown how to do this in general for surfaces in space, and that his methods readily generalised to any number of dimensions. However, the formula that Gauss had used, while natural and correct for the problem he studied, was in general too simple, and it would be necessary to consider more complicated formulae.

To see why, consider a sphere in space. We may suppose the sphere is the sphere of unit radius, and that points on it are specified by their familiar latitude ϕ and longitude θ, thus: $(\cos\theta\cos\phi, \sin\theta\cos\phi, \sin\phi)$. An increase in θ of a small amount $\delta\theta$ moves a point on the sphere on a circle of radius $\cos\phi$ through an amount $\cos\phi\, d\theta$. An increase in ϕ of a small amount $\delta\phi$ moves a point in a perpendicular direction through an amount $\delta\phi$. Following one move by another moves a point through an amount ds given by Pythagoras's Theorem as $ds^2 = d\phi^2 + \cos^2\phi\, d\theta^2$. We may, if we wish, imagine this length is measured by an infinitesimally small ruler curved to fit the surface of the sphere. Now, we know it is correct to use Pythagoras's Theorem because we are assuming we know how to measure distances in the ambient three-dimensional space, to wit, by the three-dimensional version of Pythagoras's Theorem. To find the distance between two points on a sphere we choose the shortest route between them, which we know to be an arc of a great circle joining them, we divide it up into a series of very short arcs, we estimate the lengths of these arcs using the formula just given, and we add up these arcs (which is done by integration).

Now suppose we consider just a surface, and forget entirely that it lies in space. Riemann wanted to define distances on such a two-dimensional manifold by copying what he could of Gauss's approach. He indicated the main way of doing it, which is to assume you can define distances between nearby points by a formula like the one supplied by Pythagoras's Theorem. It should involve the coordinates of the points, which are, shall we say, (x, y) and $(x + dx, y + dy)$. It will be some expression in the dx's and dy's, and the coefficients

of this expression may well involve x and y, as they did in the example of the sphere. But there is no reason to assume that the expression will resemble one derived from using Pythagoras's Theorem, so Riemann merely notices that from his perspective such formulae reveal a concealed assumption. To find the shortest distance between two points, one must then consider all paths joining them, a task Riemann did not describe how to confront.

The upshot of all this is that measuring distances in a space is something that Riemann said was part of geometry, and he indicated that one way of doing this would be to write down expressions in the coordinates and their differentials and then invoke the calculus. This is why what he outlined forms a significant shift in our ideas about differential geometry. But the underlying idea is very simple. One has a geometry whenever one has a space of points (a manifold) and a way of measuring distance between points, which would be the case if one always knew the distance between infinitesimally close points. So what one wants is a manifold and an infinitesimal ruler.

It was evident to Riemann that one might write down a vast number of different formulae for distance even given the same set of points. Presumably each formula would lead to a different geometry on that space of points, so the question was how to proceed given such a bewildering array of alternatives. There was a natural special case to consider. In the plane and on the sphere one may measure lengths by a rigid infinitesimal ruler, which may be put in any position. If the ruler is imagined to be an edge of an infinitesimal rigid body, however, then it is not true that it may be put anywhere on the surface of a pear-shaped surface, for example, because an infinitesimally small ruler adapted to fit the tightly curved dome-shaped region at the top of the pear will not fit the fatter, but still dome-shaped, region at the base, and still less is it adapted to the saddle-shaped region in the middle.

Surfaces where just one ruler will do, a ruler which one can imagine being one side of an infinitesimally small two-dimensional solid body, have particularly simple formulae for distance, and Riemann proposed to single them out. Among surfaces these include the plane and the cylinder, which cannot be distinguished when only small patches are looked at (which is why printing can be done from a cylindrical drum to a flat piece of paper) and spheres of all radii. These are among the surfaces with constant curvature, to use a term introduced by Gauss in a celebrated memoir [Gauss, 1827]. The plane and the cylinder have zero curvature. A sphere of radius R has curvature R^{-2}; so very large spheres have very nearly zero curvature, as one would expect. But it is also possible for a surface to have negative curvature. Surfaces of this kind are locally saddle-shaped. Riemann observed that if one draws triangles (whose sides are geodesics, i.e. curves of shortest length) on these surfaces and measures the sum of their angles, a simple rule emerges. On a surface of zero curvature the angle sum differs by zero from π. On a surface of constant positive curvature the angle sum differs by a positive amount from π. On a surface of constant negative curvature the angle sum differs by a negative amount from π.

But Riemann, as we have observed, wished to discuss surfaces without reference to any ambient space. His inspiration here was Gauss's discovery of curvature, which Gauss showed was something that could be determined from quantities measured in the surface alone. Indeed, Riemann's whole insight into geometry may be summarised by saying that it is about the intrinsic properties of n-dimensional manifolds, and that the study of how a manifold inherits properties from a larger ambient space should be reformulated accordingly. Now, when one studies a surface intrinsically, one only has coordinates (but not a

surface embedded in three space). As with latitude and longitude, the coordinates may be thought of as specifying points in a plane, which is a map of the surface in exactly the sense that the pages of an atlas are maps of the surface of the Earth. So, for example, to describe a sphere one might well choose the points of the plane, and measure distance according to the formula

$$ds = \frac{\sqrt{dx^2 + dy^2}}{1 + (\frac{1}{4r^2})(x^2 + y^2)}. \tag{1}$$

This corresponds to saying that the plane is the tangent plane to the sphere of radius r at the North Pole, and the point P in the plane with coordinates (x, y) corresponds to the point on the sphere which is found by joining the point P to the south pole by a straight line. Happily, starting from this formula it follows from Gauss's formula for intrinsic curvature that the space being described is indeed a space of constant curvature $1/r^2$.

This little formula for ds above conceals a remarkable statement. If one replaces $1/(4r^2y)$ by a negative quantity, say, for simplicity, -1, the formula for distance becomes

$$ds^2 = \big(dx^2 + dy^2\big)/\big(1 - \big(x^2 + y^2\big)\big), \tag{2}$$

which only makes sense when $x^2 + y^2 < 1$. But inside this region, which is the interior of the circle of radius 1, the formula for distance describes a two-dimensional space of constant negative curvature -4. There can be no doubt that Riemann knew this perfectly well, even though he did not draw attention to it. Its momentous significance will be made clear below.

At this stage in his lecture and in the paper, Riemann had outlined a programme according to which any geometry is to be thought of as a space (an n-fold extended magnitude) and distances in the space are to be measured by an infinitesimal ruler. Among these geometries, those whose distances are determined by infinitesimal n-dimensional bodies have the simplest formulae and should be studied first. He now proposed to show how this could be applied to the study of space.

Riemann first observed that, on the assumption that the infinitesimal measuring rod may be put anywhere, and so space has everywhere constant curvature, then the sum of the angles in a triangle is known once it is known in a single triangle. This recalls a famous trichotomy in the investigations of non-Euclidean geometry, discussed by most writers on that subject including (if erroneously) Legendre. If the parallel postulate is removed from the assumptions of Euclidean geometry, there are three possibilities for the sum of the angles of a triangle. Either the angle sum is always greater than π, or it is equal to π, or it is always less than π. For future reference, let us follow historical usage and call these the hypothesis of the obtuse angle, the right angle, and the acute angle respectively Further work shows that the first alternative is untenable. The second alternative is, of course, Euclidean geometry, which no one opposed. The third alternative is consistent with the remaining assumptions of Euclid, and forms the distinctive feature of non-Euclidean geometry. It is noteworthy, however, that Riemann never discussed this possibility in these terms.

Riemann next distinguished between the metrical aspects of an n-dimensional manifold and what he obscurely called the relations of extent. These may be understood as being

global, topological, or geometric in nature. The cylinder and the plane are alike metrically, but differ in their relations of extent. The point Riemann observed is that there is a distinction between a space being infinite and being unbounded. The sphere, for example, is not infinite, but it is unbounded—it has no boundary, it does not come to a stop. All the evidence is that space is unbounded, but that does not mean, said Riemann—the first time anyone had thought to say this—that it is infinite. All finite universes considered in astronomy and cosmology before had been bounded, usually by what was called the sphere of the fixed stars.

Riemann next observed that if actual space is of constant curvature, it must be of very nearly zero curvature. This is because astronomical observations have set bounds on the curvature. But if it is not of constant curvature then it is difficult to say anything about it at all. Finally, he observed that the empirical understanding of the metrical properties of space may break down in very small regions, and if this were to open the way to simpler explanations of natural phenomena we should suppose that they do. Indeed, the universe may not form a continuous (i.e. infinitely divisible) n-fold extended magnitude but may be a discrete manifold instead. He then concluded this remarkable paper with the hope that the study of space in such generality may prevent progress from being impeded by too narrow views, and would open the way to the study of phenomena otherwise difficult, if not impossible, to explain.

3 THE INTELLECTUAL CONTEXT

Most historians of mathematics have always sought to place Riemann's lecture within debates about the foundations of geometry, by which they mean the early investigations of non-Euclidean geometry, and this is indeed a reasonable thing to do. But there are other currents as well. Very few names are mentioned in the lecture. That of Gauss appears, because of his work on the differential geometry of surfaces in 1827 and some of his remarks about complex numbers, which Riemann plainly regarded as a species of two-fold extended magnitude. Legendre's name appears as someone who had not lifted the darkness lying over geometry, and that is a nod towards the study of non-Euclidean geometry because his arguments against it were all flawed, as Gauss knew and Riemann would easily have seen if he read them. The only philosopher whose name appears is that of Johann Friedrich Herbart (1796–1841). On the other hand, among the names of mathematicians which do not appear are those of Janos Bolyai (1802–1860) and N.I. Lobachevsky (1792–1856), the founders of non-Euclidean geometry, which suggests very strongly that Riemann had not read them.

Let us first deal with Herbart and Riemann's involvement with German philosophy. Herbart had briefly been a professor at Göttingen, from 1805 to 1808 and from 1833 to his death in 1841. In between he was Kant's successor in Königsberg, and his philosophy is notable for its interest in Kantian approaches at a time when German philosophy was dominated by varieties of idealism. His most important book was the two-volume *Psychology as knowledge newly founded on experience, metaphysics, and mathematics* (1824–1825). In this book he argued that philosophy was fundamental to psychology, not the other way round as had come to be suggested elsewhere. Herbart was critical of attempts to interpret Kant in psychological terms, and proposed instead a metaphysical theory of the ego and

the soul. This theory allowed him to explain, at least to his own satisfaction, how thoughts and feelings were ultimately reducible to basic presentations which led a dynamic life that generated the activity of the mind. According to this theory nothing was innate, everything was learned from experience.

The details of this dynamical system was to be understood through mathematics, and although Herbart provided few clues as to how this could actually be done, his dynamical system of the mind did suggest a novel, and much richer description of the world than the prevailing one. Herbart regarded every sensation as operating over time, and being formed of a sequence of momentary stimuli. Residues of these stimuli will be retained in the mind, and can be recalled. The perception of space, Herbart believed, form sequences that can be read in either direction. We acquire such experiences through our senses in many different ways: by our eyes, our fingers, our ears. These are woven together into the presentation of two-dimensional space; depth, Herbart believed, was an inferred quality. Time series are inevitably to be read in only one way, so our perception of time differs from our perception of space, but both series of presentations are infinitely divisible (we can imagine residua in our minds that are closer together than any residua actually present).

The influence of Herbart on Riemann can be traced through writings in Riemann's *Nachlass* that were published for the first time in Riemann's *Werke* of 1876, when the lecture was reprinted. Riemann set a high value on Herbart's work. As he put it, 'The author is a Herbartian in psychology and epistemology (methodology and eidolology); in most cases however he cannot agree with Herbart's natural philosophy and the metaphysical disciplines (ontology and synechology) referring to it' (*Works*, 540; quoted in [Scholz, 1982, 414]). Synechology covers space, time, and motion, in particular intelligible space, the mental construct that makes the explanation of matter possible. This shows that Bertrand Russell's confident ascription in 1897 of Riemann's ideas to his appreciation of Herbart was misplaced. It is not Herbart's description of space (which remained intrinsically three-dimensional) that Riemann amplified, but his ideas about epistemology and the acquisition of knowledge.

Herbart had characterised natural science as the attempt to comprehend nature by precise concepts. Such concepts determine what is probable, and if predictions based on these concepts fail the concepts must be modified. He was inclined to think that the way concepts change over was a strong argument against the Kantian categories. Riemann agreed. He wrote that it is only because concepts originate in comprehending what is given in sense-perception that '*their significance can be established in a manner adequate for natural science'* (emphasis in original: *Works*, 554). So, if all concepts arise by the transformation of earlier concepts, they need not be derived *a priori*, as were the Kantian categories. Riemann agreed with Herbart that the Kantian categories presumed too much. The idea that space was an empty vessel into which the senses ought to pour their perceptions Herbart had called 'completely shallow, meaningless and inappropriate' (quoted in [Scholz, 1982, 422]).

Riemann distinguished crucially between the space concept, and a particular space that is used to describe real events. A conception of the world is correct, he wrote, 'When the coherence of our ideas corresponds to the coherence of things', which in turn is obtained 'from the coherence of phenomena' (*Works*, 555). Herbart's philosophical programme

sought to establish philosophy as a science existing in a to-and-fro relation with the sciences in which speculation about concepts led to an ever-deepening process of education. Riemann's view of science was very similar, being heavily conceptual. As Scholz shows, Riemann's views on mathematics were deepened and clarified by his extensive studies of Herbart's philosophy. In particular, Riemann might never have formulated his profound and innovative concept of a manifold had he not immersed himself in Herbart's work.

We know from Dedekind's memoir of Riemann that when the time for the Habilitation exam came up, Riemann was working in Wilhelm Weber's physics laboratory, occupied with no less than gravitation, electricity, magnetism, and light [Dedekind, 1876]. What all these phenomena were known to have in common in 1854 was their transmission across vast distances at enormous speeds. Riemann apparently sought to explain this by imagining that the fabric of space was in some way subtly altered, and distortions in it spread like ripples. He was not able to work this up into a coherent theory, but it seems clear that this idea of conceptually rethinking the nature of space in a direct physical context accounts for many features of the lecture. It underlines the avowedly empirical nature of the lecture, it is in line with Riemann's attempt to make proposals that allow one to explain phenomena, and it explains the somewhat anti-Newtonian rhetoric. After all, the transmission and nature of gravity were problems which Newton himself had the grace to admit he could not elucidate.

There were therefore good scientific and philosophical reasons for Riemann to be thinking of geometry as the study of the metrical relations of a space, and that these should be regarded as largely arbitrary but subject to critical analysis. These reasons in turn explain why Riemann's way of thinking about the nature of space is so profoundly different from many of the ideas connected with non-Euclidean geometry, and why the link to non-Euclidean geometry is so hard to find in the lecture.

Investigators into non-Euclidean geometry in the 18th and early 19th centuries took the postulates of Euclid's *Elements* for granted, except for the parallel postulate. They then sought to deduce the parallel postulate as a theorem from the other assumptions. This heavily axiomatic approach was not Riemann's at all, and indeed he criticised it right at the start of the lecture. But it does not follow that Riemann had no interest in non-Euclidean geometry, or that he was completely unaware of what had been done on it before 1854. However, it is very hard to determine exactly what he did know.

As to what he could have known, he might have known that Gauss and his circle of intimates in the world of astronomy had been prepared for some years to entertain the idea that space departed slightly from Euclidean geometry. The remark in the lecture that the resolution of this question might exceed the current range of telescopes suggests rather strongly that Riemann was aware that the curvature of space was a topic of discussion in Göttingen. On the other hand, the fact that Riemann was surprised that Gauss chose the topic he did for the lecture suggests that Riemann was unaware of just how interesting Gauss had found the topic all his life, which might mean that the two men had not talked about it very much. The fact that the names of Bolyai and Lobachevsky do not come up in the lecture does not tell us much either, because it could simply be that Riemann had enough sense not to antagonise the philosophers who would decide if he was good enough to pass his Habilitation. It was clear at the time that if anyone could accept the views of Bolyai and Lobachevsky it was at best a few imaginative mathematicians and astronomers.

It is true also that the work of Bolyai was most obscurely published, in a little-known book [F. Bolyai, 1832]. That said, Gauss had a copy, which Bolyai's father had sent him personally; but if for whatever reason Riemann did not know it, he was unlikely to have heard of the work at all. The work of Lobachevsky was more accessible, however, for apart from his extensive Russian articles (surely known only to Gauss in Göttingen at the time) there was an article in French in Crelle's *Journal*, and a self-standing booklet written in German and published as [Lobachevsky, 1840], of which Gauss had a copy. These works did not convey to the reader the full sense in which Lobachevsky wished to rethink geometry by founding it on the idea of motion, but the booklet did at least make the new geometry plausible. However, they received little comment in the reviewing journals, and what was said was hostile (and inaccurate).

There is every reason, therefore, to doubt that Riemann ever read the work of Lobachevsky, and no reason at all to suppose he knew of Bolyai's. But there are passages in the lecture that make more sense if compared with what was generally known about investigations into the parallel postulate. One might almost claim that the only interesting thing in all Legendre's many editions of *Eléments de géométrie* (1794 and later) as its attempts on the parallel postulate. So to single him out as a terminus in investigations in the foundations of geometry from Euclid to 1854 is both fair to Legendre's revival of axiomatic approaches to geometry, as opposed to Cartesian ones, and surely a nod to his work on the parallel postulate. There Legendre had established anew G.G. Saccheri's result that if the parallel postulate is struck down there are precisely three possibilities for the angle sum of a triangle, and the angle sum of all triangles is known in all cases if it is known in one [Saccheri, 1733]. So when Riemann distinguished between the three types of constant curvature (positive, zero, and negative) on just such grounds, he was knowingly entering the topic of non-Euclidean geometry.

Riemann had no problem accepting the geometry on a surface of constant curvature as the geometry on a sphere. This is incompatible with the assumption in Euclidean geometry that all lines can be indefinitely extended, but not only was Riemann hostile to the axiomatic treatment of geometry, he was also willing to believe that space was not infinite in extent either.

Plane geometry, the geometry of a space of zero curvature, naturally posed no problems for Riemann. That left the case of a surface of constant negative curvature, and simultaneously a geometry which differs from Euclidean geometry only in assuming that the angle sum of a triangle is always less than π. The oddly tricky object here is the surface of constant negative curvature. Spherical geometry is obvious to anyone who has seen a sphere. Could there be a surface analogous to the sphere, but infinite in extent, and of constant negative curvature? The best anyone had ever found was a bugle-shaped surface, known as a pseudosphere, which [Minding, 1839] had shown had constant negative curvature. But it resembled the cylinder rather than the plane, and, which is worse, it came to an end in one direction, where it had a rim beyond which is could not be extended. This made it a very poor model for a rival to Euclidean geometry. If anything, it suggested to anyone who connected it to non-Euclidean geometry, that there might be a contradiction in the hypothesis of the acute angle after all. It is more likely, however, than no-one made the connection, because when the Italian mathematician Delfino Codazzi (1824–1873) studied the pseudosphere he established results about the trigonometry of triangles on this surface

which showed immediately that the angle sum of these triangle was less than π [Codazzi, 1857]. However, he did not notice this, and this simple observation had to wait for a further 11 years.

This is the importance of Riemann's description of a space of constant negative curvature. He did not exhibit it as a surface in three-dimensional Euclidean space. Nor need he have done: such an approach was what he was trying to get away from. He presented it as a cartographic map, defined inside a circle, and with a metric closely analogous to maps of a sphere on a plane. This presents it entirely rigorously, and it is as natural a geometry as any other in Riemann's scheme of things. What is striking is that before Riemann no-one seems to have known how to describe an unbounded, infinite surface of constant negative curvature which the study of non-Euclidean geometry seemed to require. But it must also be said that Riemann's presentation was extremely cryptic, because he simply gave a formula for the appropriate metric that can be interpreted in the way just described.

The man who put non-Euclidean geometry as traditionally defined (which means by the tradition which had generally sought to show that it could not exist) together with Riemann's ideas, and who therefore put non-Euclidean geometry securely on the map, was another Italian, Eugenio Beltrami (1835–1900). What had alerted mathematicians in the 1860s to the possible validity of non-Euclidean geometry was the discovery that Gauss had been sympathetic to the enterprise. After Gauss's death the vast treasure trove of his unpublished papers gradually began to be explored. This revealed many things that had not been known before and galvanised the community into producing an edition of his Collected Works and also of his extensive correspondence. The publication of his letters to Schumacher [Gauss, 1860–1865] showed very clearly the extent of Gauss's belief in the possibility that space might be non-Euclidean, it also drew mathematicians' attention to the original publications of Bolyai and Lobachevsky's.

4 BOLYAI AND LOBACHEVSKY AND THE DISCOVERY OF NON-EUCLIDEAN GEOMETRY

Beltrami had read [Lobachevsky, 1840] in Hoüel's French translation of 1866, having picked up the thread by reading some of what Gauss had said. In that work, and in Bolyai's Appendix of 1831, the progress of ideas is more-or-less the same (so much so in fact that historians have worked hard to establish that there was no possibility of plagiarism or of either man receiving any help from Gauss). This account therefore follows Lobachevsky's account, which is clearer and was more influential. First, a novel definition of the parallel to a given line l through a given point P in a given plane is provided. It is the line m through P such that all lines below it eventually meet l and all lines above it never meet l (Figure 1). There are two lines, one in each direction. It is fundamental to the work of Lobachevsky and Bolyai that this definition does define something new, distinct from the Euclidean definition, and that exploring the consequences of this definition will never lead to a contradiction, but this fundamental assumption is neither discussed not vindicated. Lobachevsky merely pressed on, and Bolyai noted theorems which were true whether the definition differed from the Euclidean one or not. It was therefore open to critics of their work to dismiss it as the consequences of what would surely turn out to be an untenable

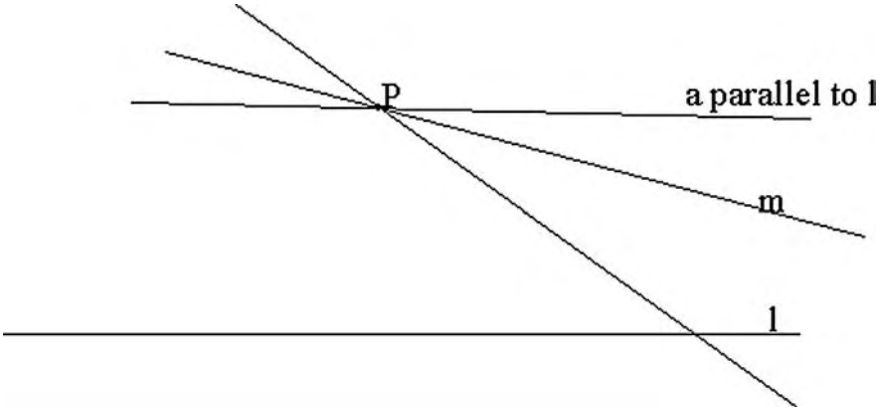


Figure 1. Lobachevsky's account of parallels.

hypothesis, but there is little evidence that the work had readers at all, let alone ones prepared to be so temperate in their judgement. The few reviewers that Lobachevsky ever had, whether in German or in Russian, simply dismissed his work as incomprehensible and full of mistakes, which suggests that they read it convinced in advance that it must be wrong.

Once the new definition of parallels is granted, theorems followed. Given a family of lines all parallel to each other (in the same direction), there is a family of curves crossing these parallels at right angles. Lobachevsky named these curves 'horocycles', a name they still retain. If a single horocycle is picked out and the whole figure rotated about one of the parallel lines, the horocycle sweeps out a bowl-shaped figure which Lobachevsky called the horosphere. Lobachevsky (and Bolyai) then proved the striking result that even if the three-dimensional space in which the bowl sits is non-Euclidean, the inherited geometry on the bowl is Euclidean geometry. So just as the sphere is a natural model of spherical geometry in Euclidean space, so the horosphere is a natural model of Euclidean geometry in non-Euclidean space. Lobachevsky then used the relation between the horosphere, a tangent plane to it, and the family of parallel lines to deduce trigonometric formulae for triangles in the (non-Euclidean) tangent plane. These formulae were closely analogous to the formulae of spherical trigonometry, which was reassuring, and they collapsed to the formulae of plane trigonometry for small triangles. This permitted the implication that small regions of space might be indistinguishable for non-Euclidean space and only astronomical regions could be shown to be different. Lobachevsky even proposed tests, but they proved inconclusive.

Lobachevsky's results should have struck fair-minded readers as plausible, if in need of sounder foundations. Bolyai had a different strategy. He cryptically outlined the way differential geometry could be done in non-Euclidean space, and challenged his readers to propose a geometric task that could not be carried out in the new setting. What is most notable, however, is that both men described a new geometry of three-dimensional space, thus showing exactly what was being said that would be a new geometry of the universe. This was something that Gauss, for all his sympathy with the idea, had never done (at least, no unambiguous evidence survives). Their work only met with general acceptance,

however, after the work of Riemann, and that of Beltrami, Felix Klein (1849–1925) and Henri Poincaré (1854–1912) to be described below. It was by then also available in French, thanks to the energy of Houël, and it was also put into English by G.B. Halsted at the end of the 19th century. Other articles by Lobachevsky originally written in Russian were translated into German at the instigation of two German mathematicians, Friedrich Engel and Paul Stäckel, who also wrote the most thorough accounts of the life of work of Bolyai and Lobachevsky [Engel and Stäckel, 1895–1913], upon which all subsequent scholarship is largely based.

5 BELTRAMI, POINCARÉ, AND KLEIN ON NON-EUCLIDEAN GEOMETRY

Beltrami's judgement on this was exactly right. He sought first to convince himself of the truth of what Lobachevsky had said, and then to find a real substrate for it, rather than to introduce new entities and concepts. His ideas on this were already well advanced before he read Riemann's lecture, but he was delayed because the eminent Italian geometer Luigi Cremona (1830–1903) was concerned that the work might contain a vicious circle. Cremona's worry was not trivial, he was concerned that ordinary Euclidean analysis was built into the fabric of the calculus, and yet the calculus was being used to establish the validity of a novel geometry. However, once Beltrami read Riemann he had the confidence to proceed. He would also have seen the description of a manifold of constant negative curvature, but he had something new to say: a detailed description of the geometry of a space of constant negative curvature, which Riemann had not discussed at all. Beltrami proceeded to show that the geometry of this space was exactly that described by Lobachevsky, thus removing any doubts about the foundations of the subject and the possibly impermissible interpretation of the definition of parallel lines [Beltrami, 1868].

Beltrami's description of non-Euclidean geometry differs in one crucial respect from that latent in Riemann's account. Beltrami decided that it would be more useful if the coordinates were chosen that curves of shortest length in non-Euclidean two-dimensional space appeared as straight lines inside the disc. He later showed that such coordinates can always be chosen if the space has constant curvature, but not otherwise. He defined a metric (an infinitesimal ruler) inside the disc by suitably modifying the metric on the sphere, and used it to deduce Lobachevsky's trigonometric formulae for triangles. From these formulae he deduced that the angle sums of triangles are always less than π. He gave formulae for the circumference and area of a circle and explained what horocycles are in the new setting. In an immediately subsequent paper he showed how to generalise all these results to n-dimensional non-Euclidean geometry. There he mentioned Riemann's name for the first time, and explicitly attributed the formula for the metric on a space of constant negative curvature to him.

Beltrami's account was soon re-interpreted by Klein, a young German mathematician on his way to becoming not, as he undoubtedly hoped in 1871, a great mathematician, but a good mathematician and a great organiser of mathematics (§42). In two papers Klein [1871, 1873] showed how Beltrami's description fitted into projective geometry, which by then was the fundamental geometry. Projective geometry is the study of the properties a figure shares with all its shadows, so being a straight line is a projective property, but being a line

segment of a certain length is not. Projective geometry was taken to be more fundamental than Euclidean geometry because any property a figure shares with its shadows, and which therefore makes no reference to the concepts of distance, is also a property in metrical, Euclidean geometry, but the converse is obviously false.

Klein's proof that non-Euclidean geometry was but a special case of projective geometry not only demonstrated the power of projective geometry, it placed non-Euclidean geometry securely in a simple geometric framework for those with little appetite for Riemann's grand vision of geometry. It was followed a decade later, in 1880, by the independent rediscovery by Poincaré of non-Euclidean geometry, which was memorably recalled by him in [Poincaré, 1908]. He recognised almost by accident that a certain complicated figure he was studying in connection with another topic entirely was identical with the Beltrami description of non-Euclidean geometry. His original problem suggested a different map, in which non-Euclidean two-dimensional space is depicted inside a disc in a way which renders angles correctly, although straight lines are now drawn as arcs of circles meeting the boundary of the disc at right angles. This account makes it easier to see the distance-preserving transformations (or congruences) of non-Euclidean geometry, which are not apparent in the projective Beltrami–Klein version. Amusingly enough, it is the Poincaré version that describes the geometry latent in Riemann's formula, although it is almost certain that Poincaré had not read Riemann's lecture at that point in his career.

6 THE LATER RECEPTION OF NON-EUCLIDEAN GEOMETRY

The work of Beltrami, Klein and finally Poincaré put an end to most mathematicians' doubts about non-Euclidean geometry, although one or two remained unsure of the new world they found themselves in. Philosophers were harder to convince, Frege being an extreme case who believed all his life that since there is only one universe there can be only one geometry (Euclidean) and that any alternative is but a fantasy [Frege, 1969]. All therefore agreed that the question of which geometry was true, Euclidean or non-Euclidean geometry, was an empirical one which, nonetheless, would inevitably be resolved by deciding that Euclidean geometry was correct for all practical purposes.

Riemann's vision of geometry had, however, been far broader, and it was to lead slowly, and not always directly, to the dominant view of the 20th century, which interprets gravity in geometric terms (compare §52). The complicated story that leads from Riemann to Einstein, Weyl and the general theory of relativity has still not been fully traced by historians of mathematics, but many of the main features are in place. A number of mathematicians wrote papers solving various of the formidable technicalities of dealing with n-dimensional geometries of variable curvature. In 1916, inspired by Einstein and Grossmann's new theory of general relativity (§63), another Italian mathematician, Tullio Levi-Civita introduced a profound idea which made it possible to discuss when a vector moves on a curve in space without changing its direction. This was the first time a significant geometric idea had been added to Riemann's theory of manifolds, and it speedily found its way into the general relativity. But this gap from 1867 to 1916, nearly 50 years, should give us pause.

In all that period there was remarkably little exploring higher-dimensional Riemannian geometry. Riemann had indicated what the constant curvature two-dimensional manifolds

are, and one might have supposed that the generalisation to three dimensions could have been proposed and energetically studied. However, the major paper in this area did not come until 1898, when it was treated by Luigi Bianchi, and even then there was not a great swirl of activity. Indeed, in the period after 1909, when Albert Einstein began to look for a way to extend the ideas of special relativity to include gravity (§63.2), there was no textbook in any language on the differential geometry of manifolds of dimension greater than 2, and almost all differential geometry was about curves and surfaces in ordinary three-dimensional Euclidean space.

BIBLIOGRAPHY

Beltrami, E. 1868. 'Saggio di interpretazione della geometria non-Euclidea', *Giornale di matematiche*, *6*, 284–312.

Bolyai, F. 1832. *Tentamen juventutem studiosam in Elementa Matheosis purae, etc.*, Maros-Vásérhely.

Bolyai, J. 1832. 'Appendix scientiam spatii absolute veram exhibens', in [F. Bolyai, 1832]. [French trans. by J. Houël, 'La science absolue de l'espace', *Mémoires de la Société des Sciences Physiques et Naturelles de Bordeaux, 5* (1867), 189–248. Italian trans. by G. Battaglini, 'Sulla scienza della spazio assolutamente vera', *Giornale di matematiche*, *6* (1868), 97–115. English trans. by G.B. Halsted, 'Science absolute of space', in [Bonola, 1912], app.]

Bonola, R. 1902. 'Index operum ad geometriam absolutam spectantium', in *Ioannis Bolyai in memoriam*, Budapest: University, 83–154.

Bonola, R. 1906. *La geometria non-Euclidea*, Bologna: Zanichelli. [English trans.: *Non-Euclidean geometry* (trans. H.S. Carslaw), Chicago: Open Court, 1912; repr. New York: Dover, 1955.]

Codazzi, D. 1857. 'Intorno alle superficie le quali hanno costante il prodotto de' due raggi di curvatura', *Annali delle scienze matematiche e fisiche*, *8*, 346–355.

Dedekind, R. 1876. 'Bernhard Riemann's Lebenslauf', in Riemann *Works*, 1st ed., 507–526. [Repr. in 2nd ed., 539–558; 3rd ed., 571–590.]

Engel, F. and Stäckel, P. 1895. *Theorie der Parallellinien von Euklid bis auf Gauss*, Leipzig: Teubner.

Engel, F. and Stäckel, P. 1913. *Urkunden zur Geschichte der Nichteuklidischen Geometrie, Wolfgang und Johann Bolyai*, Leipzig and Berlin: Teubner.

Euclid *Elements*. *The thirteen books of Euclid's* Elements (ed. and trans. T.L. Heath), 2nd ed., 3 vols., Cambridge: Cambridge University Press, 1926. [Repr. New York: Dover, 1956.]

Frege, G. 1969. *Nachgelassene Schriften*, Hamburg: Meiner.

Gauss, C.F. 1827. 'Disquisitiones generales circa superficies curvas', *Commentarii Societatis Regiae Göttingensis*, *6*, 99–146. [Repr. in *Werke*, vol. 6, 217–258.]

Gauss C.F. 1860–1865. *Briefwechsel zwischen C.F. Gauss und H.C. Schumacher*, 6 vols., Altona: Esch.

Gauss, C.F. 1900. *Werke*, vol. 8, Leipzig: Teubner.

Gray, J.J. 1986. *Ideas of space, Euclidean, non-Euclidean and relativistic*, Oxford: Oxford University Press. [2nd ed. 1989.]

Herbart, J.F. 1824–1825. *Psychologie als Wissenschaft, neu gegründet auf Erfahrung, Metaphysik und Mathematik*, Königsberg: in Commission bey A.W. Unzer. [English trans.: *A text-book in psychology*, New York: Appleton, 1896.]

Hilbert, D. 1899. *Grundlagen der Geometrie*, 1st ed. Leipzig: Teubner. [Many subsequent eds.; see §55.]

Hoüel, J. 1863. 'Essai d'une exposition rationelle des principes fondamentaux de la géométrie élémentaire', *Archiv der Mathematik und Physik*, *40*, 171–211.

Kant, I. 1781. *Kritik der reinen Vernunft*, Riga. [English trans. by N.K. Smith: *Critique of pure reason*, London: Macmillan, 1929.]

Klein, C.F. 1871. 'Über die sogenannte Nicht-Euklidische Geometrie, I', *Mathematische Annalen, 4*, 573–625. [Repr. in *Gesammelte mathematische Abhandlungen,* vol. 1, 254–305.]

Klein, C.F. 1872. *Vergleichende Betrachtungen über neuere geometrische Forschungen* (Erlanger Programm), Erlangen: Deichert. [Repr. in *ibidem*, 460–497. See §42.]

Klein C.F. 1873. 'Über die sogenannte Nicht-Euklidische Geometrie, II', *Mathematische Annalen, 6*, 112–145. [Repr. in *ibidem*, 311–343.]

Lambert, J.H. 1786. 'Theorie der Parallellinien', *Leipziger Magazin für reine und angewandte Mathematik*, 137–164, 325–358. [Repr. in [Engel and Stäckel, 1895].]

Legendre, A.M. 1794 *Éléments de géométrie*, 1st ed., Paris: Didot. [Many subsequent eds.]

Lobachevsky, N.I. 1840. *Geometrische Untersuchungen*, Berlin. [Repr. Berlin: Mayer & Müller, 1887. French trans. by J. Houël, 'Etudes géométriques sur la théorie des parallèles', *Mémoires de la Société des Sciences Physiques et Naturelles de Bordeaux, 4* (1867), 83–128; repr. Paris: Gauthier–Villars, 1866. English trans. by G.B. Halsted, 'Geometric researches in the theory of parallels', in [Bonola, 1912], app.]

Lobachevsky, N.I. 1899. *Zwei geometrische Abhandlungen* (trans. F. Engel), Leipzig: Teubner.

Minding, F. 1839. 'Wie sich entscheiden lässt, ob zwei gegebener Krummen Flächen [. . .]', *Journal für die reine und angewandte Mathematik*, *19*, 370–387.

Poincaré, H. 1908. 'L'invention mathématique', in *Science et méthode*, Paris: Flammarion, 43–63.

Riemann, B. *Works. Gesammelte mathematische Werke*, 3rd ed. (ed. R. Narasimhan), Berlin: Springer, 1990.

Saccheri, G. 1733. *Euclides ab omni Naevo vindicatus*, Milan: P.A. Montani. [English trans.: *Girolamo Saccheri's Euclides vindicatus* (ed. and trans. G.B. Halsted), Chicago: Open Court, 1920.]

Scholz, E. 1982. 'Herbart's influence on Bernhard Riemann', *Historia mathematica*, *9*, 413–440.

Stäckel, P. 1913. *Wolfgang und Johann Bolyai, Geometrische Untersuchungen, Leben und Schriften der beiden Bolyai*, Leipzig and Berlin: Teubner.

CHAPTER 40

WILLIAM THOMSON AND PETER GUTHRIE TAIT, *TREATISE ON NATURAL PHILOSOPHY*, FIRST EDITION (1867)

M. Norton Wise

This book, familiarly known as '$T\&T'$' after the authors' own habit, was the first major work in any language to transform Newtonian mechanics and Lagrangian mechanics into the new dynamics of energy of the second half of the 19th century. This reformulation derived in large part from Thomson's own earlier work: with it, extremum principles on energy functions for entire systems replaced the direct action of forces between parts; force became literally a derivative concept, the gradient of an energy function.

First publication. Oxford: Clarendon Press, 1867. xxiii + 737 pages.

Second edition. 2 vols., vol. 2 edited and with additions by G.H. Darwin. Cambridge: Cambridge University Press, 1879–1883. xxvii + 508; xxv + 527 pages.

Reprints with minor revisions. 1888–1890, 1895–1896, 1903, 1912, 1923. All Cambridge University Press.

Photoreprint of the 2nd ed. Principles of mechanics and dynamics, New York: Dover, 1962.

Abridgement. Elements of natural philosophy, Oxford: Clarendon Press, 1873. 2nd ed. Cambridge: Cambridge University Press, 1879. [Mainly the non-mathematical, 'large print' portion of the *Treatise*.]

German translation. Handbuch der theoretischen Physik (trans. H. Helmholtz and G. Wertheim), 2 vols., Braunschweig: Vieweg, 1871–1874.

Manuscripts. Section drafts, notebooks, correspondence, and other relevant materials available. See *Catalogue of the manuscript collections of Sir George Gabriel Stokes and Sir William Thomson, Baron Kelvin of Largs, in Cambridge University Library* (Cambridge, 1976) and *Index to the manuscript collection of William Thomson, Baron Kelvin, in Glasgow University Library* (Glasgow, 1977).

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

Related articles: Newton (§5), Lagrange on mechanics (§16), Fourier (§26), Green (§30), Hertz (§52), Kelvin (§58).

1 THE PLACE OF $T\&T'$ IN THOMSON'S WORK

William Thomson (1824–1907; Sir William from 1866 and Lord Kelvin from 1892) lived nearly his entire life in Glasgow. From 1845, when he became Professor of Natural Philosophy at Glasgow University at the age of twenty one, after completing his degree at Cambridge, Thomson had been pursuing solutions to physical problems in a rather unique way, developing two subjects in parallel: electricity and heat. On his life and career, including the writing of $T\&T'$, see [Smith and Wise, 1989].

Concerning electricity, to obtain the total force acting between two charged conducting spheres, Thomson employed the engineering concept of work (or 'mechanical effect') to characterize the entire system in terms of the work done to charge it, thus its work content or its total potential (soon to be potential energy). The total force was then given immediately by the rate of change of this work content for changes in the distance between the spheres, i.e., by the derivative of the potential, its gradient [Thomson, 1845, 1869]. Trivially simple in retrospect, the solution had eluded even S.D. Poisson until Thomson effectively cracked it in three lines. Both this method of solution and another based on Thomson's geometrical 'method of images' owed much to the work on electricity of George Green, whose *Essay on the mathematical analysis of electricity and magnetism* of 1828 (§30) Thomson had rediscovered in 1845. Green (and C.F. Gauss independently, whose work Thomson knew) had introduced an abstract 'potential' function for representing force as a gradient, along with a variety of theorems about potentials (for example, 'Green's theorem') that went well beyond what Gauss and Thomson had done. But it was the physical notion of the total work contained in a system that became Thomson's central concept for his continuing development of mathematical physics, and indeed for the treatment of potential theory that would enter $T\&T'$.

Thomson employed this concept of work to translate Michael Faraday's exciting researches on fields of electric and magnetic lines of force into Fourier's analytic mathematics of heat conduction. He treated Faraday's lines passing through media of varying inductive capacity as analogous to lines of heat flux passing through media of varying conductivity, so that Poisson's equation for electric and magnetic forces would have the same solutions as Fourier's continuity equation for heat conduction (§26). Using an existence and uniqueness proof for solutions to an extended continuity equation, Thomson [1848a] showed that the lines of force in the field arranged themselves so as to minimize a function that could be interpreted as the work content of the system. On this picture, a piece of soft iron near a magnet would be attracted into the field because the conducting power of the iron for lines of force tended to concentrate or focus the lines to pass through the iron, thereby reducing the work content of the field. This conduction picture, or flow analogy, with its extremum principle governing the entire field, supplied the first mathematical alternative to the action-at-a-distance picture of P.S. Laplace, Poisson, Gauss, and Green. It reified work content as the physical entity that constituted the field, providing a mathematical foundation for what he would call 'mechanical energy' in 1851, with work as its

measure. And it was on this mathematico-physical foundation that Clerk Maxwell would build his first theory of electric and magnetic fields in 1855, 'On Faraday's lines of force' [Darrigol, 2000, 113–136].

The second major path that Thomson trod concerned the nature of heat and of the work done by heat engines. He and his engineering brother James became involved with the issues from the early 1840s, when they learned of Sadi Carnot's work of 1824 on *The motive power of fire* (through an English translation of Emile Clapeyron's mathematical treatment of 1837, which introduced the 'Carnot diagram'). Only in Paris in 1845 was Thomson able to locate the original, a rediscovery as important for energy physics as that of Green's *Essay*. Carnot had understood the steam engine by analogy to a waterfall driving a waterwheel. Just as a weight of water falling through a height produced the work done by the wheel, so a quantity of heat 'falling' from high to low temperature produced the work of the engine. Thomson used this scheme to produce the first version of his absolute scale of temperature, according to which one degree is the temperature difference between the high and low temperature reservoirs of a reversible heat engine, or 'Carnot engine', if the fall of one unit of heat produces one unit of work [Thomson, 1848b].

This basic scheme for defining absolute units in terms of work would continue in the ultimate Kelvin scale of temperature, but its original version was flawed from the beginning because Thomson had already heard James Joule in 1847 present his experiments showing that work could be converted into heat, simply by stirring water, which raised its temperature. Joule's measurements showed that the work done was linearly proportional to the temperature rise and therefore to the heat produced, rather than proportional to the square of the heat produced, as it should have been on a reversed waterfall analogy, in which the temperature difference would be produced by doing work to raise the heat. According to Joule, the work done by an engine derived not from the fall of heat but from the conversion of heat. For in his view heat was itself nothing other than the motion of molecules, or *vis viva*, produced by work and convertible back into work as part of the universal conservation of *vis viva*.

Although Thomson shared with Joule the general belief in conservation of work content, it would take him three years, working with Joule, his brother James, and fellow Scot W.J.M. Rankine, to reconcile that belief with the apparently inevitable losses that attended every fall of heat, the operation of all real engines, and, indeed, all physical processes whatever. The familiar result was his dynamical theory of heat of 1850–1851 and the two laws of thermodynamics, asserting that mechanical energy is always conserved in physical processes and equally that mechanical energy is always being dissipated, or lost to mankind for the production of work, in those same processes [Thomson, 1850–1851]. 'Mechanical energy' now explicitly consisted of two forms, 'statical' and 'dynamical', which at Rankine's suggestion became 'potential' and 'actual' in 1852 and then 'potential' and 'kinetic' in 1862, during the writing of $T\&T'$.

These two major developments, field theory and thermodynamics, changed the foundations of mathematical physics. Together with the simultaneous but largely independent work of Hermann von Helmholtz on conservation of force (1847) and Rudolf Clausius on thermodynamics (1850), they established the primacy of work content (energy) in physics. But more was required. The molecular motions constituting heat in the dynamical theory remained undefined; likewise the fields of electricity and magnetism required a mechanical

basis, presumably in the luminiferous ether that was supposed to ground the wave theory of light and radiant heat. And this ether had to interact intimately with normal matter in order to explain how molecular motions could radiate waves of heat and light into the ether.

From the early 1850s Thomson sought a solution to all of these problems in a substratum conceived as a continuous, frictionless fluid or 'aer' (a-eth-er) whose motions and distributions would explain the properties of both ether and matter. Heat, light, and electromagnetism would all find a dynamical explanation, as would interactive phenomena like the magnetic rotation of polarized light and thermo-electricity, together with speculative possibilities like thermo-magnetism and thermo-elasticity [Knudsen, 1976]. For much of the rest of his long life, Thomson avidly pursued this unifying dynamical theory. Most prominent among his attempts were the theory of vortex atoms and the vortical structure of an elastic ether. Hydrodynamics was to be the foundation of all physical science. But a major gap remained. Not only hydrodynamics, but mechanics in general, had not yet received a systematic reformulation commensurate with the physics of work and energy. $T\&T'$ would take up that task.

2 COLLABORATION WITH TAIT

Thomson's collaboration with Peter Guthrie Tait (1831–1901) began in 1861 after Tait returned to Edinburgh to become professor of natural philosophy. Like Thomson in Glasgow, Tait lived virtually his whole life in Edinburgh, aside from his mathematical training at Cambridge and four years as professor of mathematics at Belfast (where Thomson's brother James was professor of engineering). While Thomson was already famous in both scientific and broader circles when the collaboration began, Tait was just beginning his career, having published only nine papers since his degree at Cambridge in 1852 with the highest mathematical honors obtainable. He had begun work on an elementary textbook suitable for his new course on natural philosophy, and Thomson agreed to join the project. The lack of such a book was widely felt. With enunciation from many quarters of ideas of conservation of 'energy', whether as powers, forces, work, duty, mechanical effect, mechanical value, or labouring force [Kuhn, 1959], the need to restructure how natural philosophy was conceived and taught had often been expressed. Thus Thomson and Tait hoped, as Tait put it, to 'astonish the world with [. . .] what it has not yet seen, a complete course of Natural Philosophy, Expl & Mathematical' [Smith and Wise, 1989, 352].

The original plan called for three volumes, the first two experimental, the third mathematical. The experimental volumes, fairly straightforward in content, would support a descriptive course of lectures with demonstration experiments. The mathematical volume, the unique one, Tait predicted 'would go over Europe like a statical charge'. Gradually, however, the mathematical part swallowed the experimental as Thomson continually inserted mathematical notes (small print) into the descriptive discussions (large print). This printing convention mirrored the division of his own natural philosophy course, with separate hours for popular demonstration lectures and for more advanced mathematical treatment. As the small print engulfed the large, the first volume, intended to be short and popular, became over 700 pages of kinematics and abstract mechanics, missing even planned chapters on properties of matter, sound, and light. The second volume, intended to cover heat, magnetism, electricity, and electrodynamics and to 'finish up with a great section on the *one*

law of the Universe, the Conservation of Energy', never saw print of any size. Nevertheless, $T\&T'$ would make this one law the basis for all of mechanics, and indeed for all of natural philosophy, conceived as ultimately dynamical [Knott, 1911, 176–182].

The process through which the book actually got written over six years tells a great deal about the experience of reading it. Tait was to be responsible for most of the drafting with Thomson contributing ideas, revisions, expansions, and notes that Tait could incorporate. An orderly exchange should have taken place through such means as notebooks mailed back and forth between Glasgow and Edinburgh and then through joint correction of the proofs, an arrangement which should have produced the three short volumes in little more than a year. Tait might have managed it, even though notebooks sometimes got mislaid. Thomson, however, could never accommodate himself to such order. He sent scraps of this and that, a letter that was half manuscript on absolute units and half railing about papists, or nothing at all for months while he was engaged with the Atlantic telegraph or traveling with his very ill wife to health spas on the Continent. Worse, when he did return corrected proofs, he sometimes added more on a sheet than it had originally contained.

The resulting text was (and is) sometimes difficult to follow and not always rigorous; its contents is summarized in Table 1. But it was also full of insights, new formulations, and suggestive directions for the new physics. Two overriding characteristics are apparent. First, atoms, in the sense of discrete hard balls, nowhere appear, for Thomson was by then dead set against such entities and already pursuing the continuum theory of vortex atoms and the ether, built on 'the hypothesis that space is continuously occupied by an incompressible frictionless liquid acted on by no force [sic], and that material phenomena of every kind depend solely on motions created in this liquid' [Thomson, 1869]. Force was to be explained by underlying motions. The new dynamics, therefore, could not be a reductive mechanics of atoms and forces, in the tradition of Newton, Laplace, Helmholtz and most contemporary continental theorists. Secondly, at every opportunity, $T\&T'$ insisted on a practical, common-sense understanding of the mathematics they employed. Their mechanics, even when abstractly formulated, was a mechanics grounded in machinery and engineering.

3 KINEMATICS

Appropriate to the ultimate goal of establishing a physics of motions, and in agreement with modern French and British usage, $T\&T'$ separated the purely geometrical science of motion, *Kinematics*, from considerations of matter and force, which they called *Dynamics* and divided into statics and kinetics (thus kinetic energy, kinetic friction, etc.). Kinematics derived from the French engineering tradition from Lazare Carnot, G. Monge, and A.-M. Ampère to the contemporary writers on mechanics for the *Ecole Polytechnique* such as J.M.C. Duhamel and C. Delaunay. It had been adopted at Cambridge in 1841 in a complementary pairing of textbooks on machines that separated the *Principles of mechanism* (geometrical) by Robert Willis from *The Mechanics of engineering* (causal) by William Whewell. Willis's pure mechanism, or kinematics, concerned the changes of motion allowed by the various possible connections between parts of machines (for example, rod and crank, or rack and pinion), yielding a taxonomy of joints and motions. The analysis

Table 1. Summary by Sections of the *Treatise on natural philosophy*.

Page	Section contents
v	Preface; Natural philosophy defined; survey of Vol. I (no more published)
	Division I. Preliminary notions
	Chapter I. KINEMATICS (160 pages)
1	Geometry of motion; velocity and acceleration, angular motion, relative motion.
36	Simple harmonic motions, composite motions; Fourier analysis.
56	Curves produced by mechanism; analysis of curvature; bending.
98	Analysis of strain: strain ellipsoid, shear, heterogeneous strain, non-rotational strain.
124	Equation of continuity; freedom and constraint; generalized coordinates.
137	Appendices: Green's theorem extended; Laplace's spherical harmonics.
	Chapter II. DYNAMICAL LAWS AND PRINCIPLES (144 pages)
161	Definitions of dynamical terms; Newton's laws of motion; D'Alembert's principle.
187	Energy: Newton; conservation; virtual velocities as energy principle; least constraint.
206	Impulsive motion and extremum theorems on energy.
231	Hamiltonian dynamics: principles of least action and varying action; equations of motion; characteristic equation; use in optics.
251	Lagrangian dynamics: equations of motion; Hamilton's form; examples.
270	Disturbance of equilibrium; dissipative systems.
282	Kinetic stability: hydrodynamical; projectile; principle of varying action and examples.
305	Chapter III. EXPERIENCE (16 pages)
	Observation and experiment; hypotheses; least squares.
321	Chapter IV. MEASURES AND INSTRUMENTS (16 pages)
	Time, space, mass, force, and work; their instruments.

extended to the kinds of curves that could be produced by mechanism, such as epicycloids and conchoids. Thus kinematics, considered as geometry, embodied the view that a curve should be understood in terms of the process of generating it by the motion of a point, rather than simply as a static object described by an algebraic equation. Similarly, surfaces were generated by the motion of a line. This view correlated well with Isaac Newton's fluxional calculus as opposed to J.L. Lagrange's abstract algebraic calculus, a happy circumstance for Thomson and Tait.

A good example of the kinematics of $T\&T'$ is their presentation of a non-reentrant hypotrochoid (see Figure 1) as the composition of two circular motions. They note that it is 'of very great importance in modern physics' with respect to the rotation of the plane of polarized light and that it is 'the path of a pendulum-bob which contains a gyroscope in rapid rotation' (pp. 50–51). Thomson's ether vortices lie in the background. More generally, the geometrical form, the mechanism to produce it, and the physical phenomenon are closely interrelated. This example is the last in a 20-page section on simple harmonic motions. It concludes with five pages of small print on Fourier analysis, of whose importance

Table 1. (*Continued*)

Page	Section contents
	Division II. Abstract dynamics
	Chapter V. INTRODUCTORY (5 pages)
337	Abstract dynamics: perfectly rigid or elastic solids; frictionless, incompressible fluids.
	Chapter VI. STATICS OF A PARTICLE–ATTRACTION (70 pages)
342	Equilibrium of forces, geometrical theorems for spherical shells.
363	Potential energy, potential, Laplace and Poisson's equations, equipotential surfaces.
373	Surface theorems and Green's problem. Method of images.
388	Ellipsoidal bodies; spherical harmonics; Green's method as energy formulation.
	Chapter VII. STATICS OF SOLIDS AND FLUIDS (316 pages)
412	Equilibrium of a rigid body: forces and couples.
427	Flexible cord; elastic wire: torsion and flexure, spiral springs, planks, hoops.
475	Elastic plate: synclastic & anticlastic stress; potential energy; equation of bent surface.
506	Elastic solid: stress and strain; potential energy; 21 coefficients of elasticity; equations of equilibrium; torsion of prisms (Saint-Venant); spheres and spherical harmonics.
590	Hydrostatics. Equilibrium: floating body, ellipsoid of revolution; rotating ellipsoid.
618	Digression on spherical, polar, zonal, etc. harmonics, with figures and tables.
633	Digression on potential theory with respect to figure of Earth and sea.
649	Corrected equilibrium theory of the tides.
669	Figure of the rotating Earth: as heterogeneous liquid; Laplace's interior density hypothesis; precession and nutation; abrupt changes of density.
689	Rigidity of the Earth: tides in an elastic solid; effect of solid tides on liquid tides; conclusions: Earth more rigid than glass; little fluid in the Earth.
705	Appendix: equations of equilibrium of elastic solid deduced from principle of energy.
711	Appendix: secular cooling of the Earth. [End 727.]

the phenomena of sound waves, telegraph signals, and the cooling of the Earth (some of Thomson's favorite subjects) are said to provide only a feeble idea.

After learning a great deal about curves rolling on surfaces and a variety of geometrical theorems from Leonhard Euler and Gauss, the diligent student would come to another subject crucial to any treatment of the ethereal continuum, the description of strains in solids and liquids. Adopting the macroscopic approach of their close friends Stokes and Maxwell, as opposed to the microscopic (atoms and forces) treatments of Cauchy and other continental mathematicians, $T\&T'$ led their readers through strains produced by dilation, shear, and rotation, along with the mathematical techniques for describing them in terms of the strain ellipsoid, principle axes, and Green's 21 coefficients of what would today be a strain tensor, all carried out in the lengthy form of Cartesian component equations.

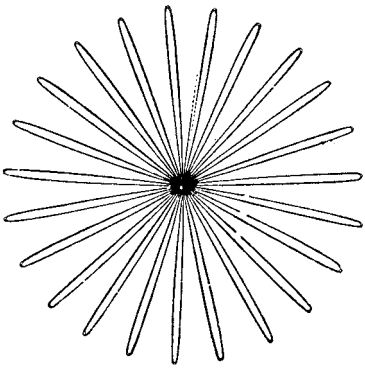


Figure 1. A hypotrochoid, that is, the path of a pendulum-bob containing a gyroscope in rapid rotation (p. 50).

In short, in their first chapter of 160 pages on kinematics, $T\&T'$ provided a compendium of the latest mathematical techniques for describing motions, concluding with appendices on Green's theorem and on the important spherical harmonics of Laplace. What they did not include, and the source of a 'thirty-eight year war' between the two authors, was the quaternion algebra of Sir William Rowan Hamilton, on which vector analysis was soon to be based (§35.5). By the time their collaboration began, Tait had already published three articles on quaternion investigations of Jean Fresnel's wave surface and of electrodynamics and magnetism and would effectively become the heir of quaternions after Hamilton's death in 1866. His *Elementary treatise on quaternions* appeared in 1867. He expected to use them throughout $T\&T'$ as an instrument of simplification and mathematical insight; but to Thomson they came wrapped in Hamilton's Idealist metaphysics, and their symbolic form obscured the object of the analyst's attention. It was a blind spot that many later analysts found puzzling at best [Knott, 1911, 119–175, 185].

4 DYNAMICS OF ENERGY

A closely related curiosity is that although Hamilton had reformulated mechanics in 1834–1835 on the basis of extremum principles, and although he was well known to natural philosophers (physicists) as a mathematician, his approach was not taken up in mathematical physics until it appeared in $T\&T'$, after Tait showed Thomson its relevance. The reasons for the previous long delay are surely to be sought in the fact that the extremum principles had not been seen as operating on anything physically real, but provided only abstract methods for deriving some new and physically uninteresting equations of motion to replace the perfectly functional equations obtained by Lagrange. In short, the mathematics waited for the physics of energy to provide the missing reality.

To appreciate the task that Thomson and Tait faced in rewriting mechanics as the dynamics of energy, it is helpful to recall that no agreement existed on the proper foundations of mechanics. Newton's three laws of motion were not even cited in French works and

while British texts did give three laws they were usually not Newton's. The relation of statics (causes of equilibrium) to dynamics (causes of motion) was equally problematic. The great Lagrange, in his *Méchanique analytique* (1788), had begun from the generalization of Newton's third law for a many-body system known as 'd'Alembert's principle', which states that the applied forces must be in equilibrium with the reversed effective forces, that is, the reversed accelerating forces or reaction forces (§11). Putting this statement into Jean Bernoulli's general condition of equilibrium for static systems, the variational principle of virtual velocities, Lagrange reduced all problems of dynamics to statics, and he derived from this condition the 'Lagrangian' equations of motion. Interestingly, some important French textbooks, such as that of Poisson, followed his reduction of dynamics to statics but did not employ his equations of motion.

In Britain, Cambridge textbooks by Whewell, J.H. Pratt, and S. Earnshaw dismissed the entire Lagrangian apparatus, beginning by rejecting the view that static forces were identical with forces producing motion and, in Whewell's case, continuing with a rejection of the principle of virtual velocities, even for equilibrium. In Scotland the French had a much stronger following, so that Thomson and Tait could follow Scottish tradition in treating dynamics and statics on the same basis (motion produced and motion destroyed). But Lagrange's reduction of dynamics to statics violated their ultimate goal of explaining force by underlying motions. They aimed rather at treating statics as a special case of dynamics, the dynamics of energy.

To work their energy revolution, then, Thomson and Tait would go back beyond Lagrange to seek authority in the unrivaled hero of British science, Newton. In reading his *Mathematical principles of natural philosophy*, they discovered that he had interpreted his third law of motion, the action-reaction law, in a way that would allow them to read it not as an action and reaction of forces, or an exchange of momenta, but as an exchange of energies. The key to this reading was the view Thomson had been developing for years, that force should be understood simply as a convenient expression for the rate of doing work. Newton's third law, on this understanding, was just the law of conservation of energy. From it, they believed they could derive the whole of dynamics, including statics. They would reverse Lagrange while restoring Newton, fashioning their *Treatise* as the *Mathematical principles* of the modern era, through the energy principle (p. 200):

> The whole work done in any time, on any limited material system, by applied forces, is equal to the whole effect in the forms of potential and kinetic energy produced in the system, together with the work lost in friction. This principle may be regarded as comprehending the whole of abstract dynamics, because [. . .] the conditions of equilibrium and of motion, in every possible case, may be immediately derived from it.

Here work lost by friction is not actually lost but produces the motions of heat. Lacking any specific knowledge of these 'inscrutably minute motions'—or those of electricity, magnetism, light, or chemical affinity—they included sliding friction as a place-holder for them until the 'universally conservative character of all dynamic action' could be scrutinized more thoroughly (p. 195).

5 EXTREMUM PRINCIPLES

While the grand claim to be able to derive the equations of motion from the energy principle alone conveys Thomson and Tait's enthusiasm for a new dynamics, it was overextended. To be valid, as they in fact proceeded to show, it required that the conservation principle be treated as a variational principle, equivalent (for conservative systems) to Lagrange's principle of virtual velocities, which considers all possible infinitesimal displacements that the system might be made to undergo and locates the actual motion where the variation vanishes. This variational condition places an additional requirement on the conservation principle, yielding the equations of motion. In its variational form the energy principle led quite naturally to a whole series of maximum/minimum theorems on what was now the kinetic energy of mechanical systems, theorems that had been established by Euler, Lagrange, and others up to the mid-19th century. These theorems, including extensions by Thomson himself, had always been closely related to the extremum principle that $T\&T'$ would now use to complete the dynamics of energy and to gain access to Hamilton, namely, the principle of least action.

It was Tait who finally sat Thomson down and showed him in three simple pages (preserved in Thomson's notebook) that Hamilton's procedures were all about energy and that they followed naturally from least action. The principle of least action, developed by P.L. Maupertuis and Euler in the mid 18th century (§16.3), had been rejected by Lagrange and Laplace as too metaphysical and was said by $T\&T'$ to have been 'regarded rather as a curious and somewhat perplexing property of motion, than as a useful guide in kinetic investigations' (p. 231). It would now gain a perfectly straightforward meaning: the time integral of the kinetic energy of any conservative mechanical system over its natural path from one configuration to another (its 'action') must be a minimum, or its average kinetic energy multiplied by the time must be minimum. In variational terms this meant that the variation of the action integral over all paths that the system might be guided to take between the same end points had to vanish, from which followed immediately the variational form of the energy principle, and therefore the Lagrangian equations of motion (p. 253).

Hamilton had converted these equations into a form that could now be seen to directly manifest the role of energy. The Lagrangian equations are second-order partial differential equations. In his system of generalized coordinates, there are n equations, one for each of the n degrees of freedom in the system, and the coordinates and their corresponding momenta are not regarded as independent. The Hamiltonian equations consist of $2n$ first-order equations given in pairs, in which the generalized coordinates and momenta are regarded as independent and play symmetrical roles. For a conservative system, the time rate of change of any coordinate q_i (or momentum p_i) is given by the positive (or negative) partial derivative of the energy E with respect to the corresponding momentum (or coordinate) (p. 254):

$$dq_i/dt = \partial E/\partial p_i \quad \text{and} \quad dp_i/dt = -\partial E/\partial q_i. \tag{1}$$

These elegant 'canonical equations' clearly made energy the central entity on which hung the description of the entire system.

Hamilton had passed from the principle of least action to a related principle with quite different properties, the principle of varying action. While the variation involved in least

action concerns virtual, or guided paths near the natural path with the same end points and the same constant energy; varying action considers only natural paths but with varying end points and varying energy. Hamilton's procedure yielded a differential equation of the first order and second degree governing what he called the 'characteristic function'. It is again the action but expressed in terms only of the end points and the energy. To solve the equation would usually be difficult, but in principle, knowledge of the characteristic function A would immediately give the momenta and the time at the end points by simple differentiation (p. 236):

$$p_i = \partial A/\partial q_i \quad \text{and} \quad t = \partial A/\partial E. \tag{2}$$

Repeatedly calling this result 'remarkable', $T\&T'$ noted that Hamilton had made great use of it for the three-body problem in planetary astronomy and for a general theory of optical instruments, or geometrical optics, to which they hoped to return in a later volume. By the time they wrote, a number of other mathematicians had long been working with and extending the Hamiltonian variational schemes and their applications, most famously C.J.G. Jacobi, but also J. Liouville, Arthur Cayley, and George Boole, among others. In their second edition (1879) the Lagrangian and Hamiltonian affairs were thoroughly revised, cleaned up, and extended to include such things as 'ignoration of coordinates'. This technique was developed by the Cambridge mathematician E.J. Routh for eliminating from the explicit description of a system in Lagrangian or Hamiltonian terms any coordinate corresponding to a constant momentum. For example, the gyroscopic motion that was supposed hidden inside the pendulum bob that produced Figure 1, could be treated simply as a constant of the motion and ignored in the equations of motion. The technique was crucial for obtaining external macroscopic descriptions of systems like the ether whose internal motions were unknown.

6 ABSTRACT DYNAMICS

With their principles in place, Thomson and Tait launched into their comprehensive treatment of all natural philosophy by attacking the part that would allow them to deal only with the first law of energy, its conservation, and not the second, its inevitable dissipation. They would treat initially only the equilibrium dynamics, or statics, of abstract matter, meaning perfectly rigid or perfectly elastic solids and frictionless, incompressible fluids. And again, only macroscopic characteristics would enter, consistent with the view that the dynamics of a continuum would ultimately explain all physical phenomena. They aimed, first, to teach physically useful methods of mathematical analysis and, second, to show how traditional formulations could (and should) be expressed in terms of energy. Proceeding in the geometrical order of point, line, surface, and body, they would take up particles, cords and wires, plates, and solids and fluids.

The statics of a particle gave Thomson and Tait the occasion to survey the theory of potentials for the inverse-square forces of gravity, electrostatics, and magnetostatics. Their discussion goes little further than the work of Green, Gauss, and Thomson from the 1840s. It gives, however, a concise discussion of Green's potential theory and surface theorems,

the method of images, the attraction of ellipsoids, and of the application of spherical harmonic analysis. Of course $T\&T'$ made the mathematical theory of potentials into the physical theory of potential energy.

The same thing should be said of the entire statics. Thus the equilibrium of a wire bent and twisted into a spiral spring and slightly stretched, could be described by two couples of flexure and torsion, and these couples could be represented simply by differential coefficients of the potential energy of the spring (pp. 446–451). Of course this would have to be true in general. The great value of $T\&T'$ lay in their showing how to express the energy for such systems and to relate it to the stresses. Their statics of solids and fluids is full of such mathematical techniques and corresponding theorems, which it will not be possible to survey here.

Of particular interest, however, both mathematically and physically is the long section on hydrostatics as applied to the figure of the Earth and the sea, including tides in the sea and in the Earth. The most extensive previous work on these subjects had been done by Laplace (§18.5). $T\&T'$ made important corrections to Laplace's famous equilibrium theory of the lunar and solar tides (based on water covering a spherical solid earth with no interruptions of land) to arrive at a more accurate theory. More interestingly, they extended his work on how the figure of the Earth could be obtained from the equilibrium of rotating spheroidal shells of originally molten matter, matter that increased in density with depth and pressure inside the Earth. Considerations of precession, nutation, tides in the sea, and tides in the Earth led them to a striking conclusion: the Earth must be more rigid than steel and contain little fluid (pp. 689–690).

For the second edition of 1879–1883, their friend and editor of the abstract dynamics, George Darwin (1845–1912), a son of Charles, would make extensive additions and corrections to this entire hydrodynamical theory without, however, changing the conclusion. His appended analysis of the retarding effects of tidal friction was incorporated even into H.G. Wells, *The time machine* (1895).

7 RECEPTION

Thomson and Tait's long-awaited *Treatise on natural philosophy* never progressed beyond the abstract dynamics of systems in equilibrium and contained nothing of the new physical theories that motivated it in the first place, namely, electromagnetism and thermodynamics; nevertheless, it immediately became a defining textbook for the physics of energy. Widely reviewed and praised in Britain, it was quickly translated by the German founder of energy physics, Hermann Helmholtz (with G. Wertheim), as a 'handbook' of theoretical physics, meaning not only a manual but also a guidebook. As Helmholtz wrote in his preface with respect to Thomson's penetrating intellect, the book leads the reader 'into the workshop of his thoughts' where with the aid of his gifted collaborator he sorted out the 'tangled and intractable material' in the new analytic framework.

It was perhaps at Cambridge University, the center of mathematical education in Britain, that $T\&T'$ played its most important role in establishing a new practice of mathematical physics, especially when joined with Maxwell's *Treatise on electricity and magnetism* of 1873 (§44) and Lord Rayleigh's *Theory of sound* (1877–1878) (§45). Together, the

three textbooks defined 'dynamical theory'. Based on energy and articulated mathematically through Lagrangian and Hamiltonian techniques, dynamical theory aimed to describe physical processes at the level of experimentally accessible parameters independent of specific hypotheses about underlying physical reality, but always on the assumption that the reality consisted in motion. It constituted the program of Cambridge mathematical physics for the remainder of the century [Warwick, 2003, 324; Buchwald, 1985, 225–228; Wise, 1982].

BIBLIOGRAPHY

Buchwald, J. 1985. 'Modifying the continuum: methods of Maxwellian electrodynamics', in P.M. Harman (ed.), *Wranglers and physicists: studies on Cambridge physics in the nineteenth century*, Manchester: Manchester University Press, 225–241.

Darrigol, O. 2000. *Electrodynamics from Ampère to Einstein*, Oxford: Oxford University Press.

Knott, C.G. 1911. *Life and scientific work of Peter Guthrie Tait*, Cambridge: Cambridge University Press.

Knudsen, O. 1976. 'The Faraday effect and physical theory, 1845–1873', *Archive for history of exact sciences, 15*, 235–281.

Kuhn, T.S. 1959. 'Energy conservation as an example of simultaneous discovery', in M. Clagett (ed.), *Critical problems in the history of science*, Madison: University of Wisconsin Press, 321–356. [Repr. in *The essential tension*, Chicago: University of Chicago Press, 1977, 66–104.]

Smith, C. and Wise, M.N. 1989. *Energy and empire: a biographical study of Lord Kelvin*, Cambridge: Cambridge University Press.

Thomson, W. *Papers. Mathematical and physical papers*, 6 vols., Cambridge: Cambridge University Press.

Thomson, W. 1845. 'Note sur les lois élémentaires d'électricité statique', *Journal des mathématiques pures et appliquées, (1) 10*, 209–221. [English version, much extended, in [Thomson, 1872], 15–37.]

Thomson, W. 1848a. 'Theorems with reference to the solution of certain partial differential equations', *Cambridge and Dublin mathematical journal, 3*, 84–87. [Repr. in [Thomson, 1872], 139–143.]

Thomson, W. 1848b. 'On an absolute thermometric scale, founded on Carnot's theory of the motive power of heat', *Philosophical magazine, (3) 33*, 313–317. [Repr. in *Papers*, vol. 1, 100–106.]

Thomson, W. 1851–1853. 'On the dynamical theory of heat, with numerical results deduced from Mr. Joule's "Equivalent of a thermal unit" and Mr. Regnault's "Observations on steam" ', *Proceedings of the Royal Society of Edinburgh, 3*, 48–52; full paper in *Transactions of the Royal Society of Edinburgh, 20*, 261–288, 475–482. [Repr. with other work in *Papers*, vol. 1, 177–291.]

Thomson, W. 1853. 'On the mutual attraction or repulsion between two electrified spherical conductors', *Philosophical magazine, (4) 5*, 287–297; *(4) 6*, 114–115. [Repr. in [Thomson, 1872], 86–97.]

Thomson, W. 1869. 'On vortex motion', *Transactions of the Royal Society of Edinburgh, 25*, 217–260. [Read 1867. Repr. in *Papers*, vol. 4, 13–66.]

Thomson, W. 1872. *Reprint of papers on electrostatics and magnetism*, London: Macmillan. [2nd ed. 1884.]

Warwick, A. 2003. *Masters of theory: Cambridge and the rise of mathematical physics*, Chicago: University of Chicago Press.

Wise, M.N. 1982. 'The Maxwell literature and British dynamical theory', *Historical studies in the physical sciences, 13*, 175–205.

CHAPTER 41

STANLEY JEVONS, *THE THEORY OF POLITICAL ECONOMY*, FIRST EDITION (1871)

Jean-Pierre Potier and Jan van Daal

Jevons's *Theory* was the first book in which economics is presented as a mathematical science based on the assumption of utility-maximising individuals. The following questions are dealt with: theories of pleasure and pain, utility, exchange, labour, rent and capital. Jevons's main contribution, however, concerns the theories of utility and exchange; these are the subjects of this article.

First publication. London: Macmillan, 1871. xvi + 267 pages.

Photoreprint. Ed. and intro. by Bert Mosselmans and Michael White; London: Palgrave, 2001.

Later editions. 2nd ed. revised and enlarged, with new preface and appendices, 1879. 3rd ed. 1888 (preface by Harriet Ann Jevons). 4th ed. 1911 (preface by Herbert Stanley Jevons), repr. 1957, 1965. All London: Macmillan; some also with New York: Augustus M. Kelley.

Other edition. Ed. and intro. by R.D. Collison Black, Harmondsworth: Pelican Classics, 1970.

French translations. 1) *La théorie de l'économie politique* (of 3rd ed., trans. H.-E. Barrault and M. Alfassa), Paris: Giard et Brière, 1909. 2) Manuscript of a French trans., *Théorie de l'économie politique*, by Léon Walras (January 1880), based on the 2nd ed.; 6 notebooks, preserved in the Fonds Walras, *Bibliothèque Cantonale et Universitaire de Lausanne*; to be published by J.-P. Potier and J. Van Daal.

German translation. *Die Theorie der politischen Ökonomie* (trans. O. Weinberger), Jena: G. Fischer, 1923.

Italian translations. 1) Of 1st ed. by G. Boccardo, as 'La teorica dell'economia politica', *Biblioteca dell'Economista*, ser. 3, vol. 2, Turin: Unione Tipografica Editrice, 1875, 173–311. 2) Of 3rd ed. by R. Fubini, in *Teoria della economia politica ed altri scritti*

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

economici (intro. by L. Amoroso), Turin: Unione Tipografica Editrice Torinese, 1947, 1–221.

Spanish translation. *La teoría de la economía politica* (intro. M.J. González, trans. J. Pérez-Campanero, rev. C. Rodrígez Braun), Madrid: Pirámide, 1998.

Related article: Shewhart (§31).

1 A NEW THEORY OF VALUE

Stanley Jevons (1835–1882) passed a busy and unusual career in his short life. While best remembered for his work in economics [Schabas, 1990; Peart, 1996], he made notable contributions also to aspects of physics, logic, and the philosophy of science [Jevons, 1874]. After some early years in Australia, he studied at University College London and was Professor of Political Economy there from 1876 after holding posts at colleges in Manchester and Liverpool.

Jevons's *Theory of political economy* is the result of research continued after his presentation in October 1862 before the Section F ('Economic Science and Statistics') of the British Association for the Advancement of Science at Cambridge of his 'Notice of a general mathematical theory of political economy' [Jevons, 1866]. In spite of its title, this paper did not contain any formulae, but the germs of his more explicit mathematical results were clearly present already [Grattan-Guinness, 2002].

The contents of Jevons's book are summarised in Table 1. He was explicitly inspired by Jeremy Bentham's so-called 'utilitarianism' [Bentham, 1789], which can roughly be summarised in the following three (not wholly coherent) principles: i) During his lifetime, Man maximises his utility (or, if one so wishes, happiness), which depends on 'his pleasures and pains'; ii) Man's individual behaviour must be based on good instruction and adequate legislation and iii) The ultimate goal of society is maximisation of total happiness of all people together. Jevons went even so far as to declare (*Theory*, 44): 'Pleasure and pain are undoubtedly the ultimate objects of the Calculus of Economy'. Among the different 'circumstances' relating to pleasure and pain envisaged by Bentham, Jevons selected in particular their intensity and duration as subjects for his analysis (without much further clarification, however) and he passed without any comment from the duration of the pleasure of some good to its quantity consumed.

Further, it must be observed that in 1871 Jevons was much more sceptical about the possibility to measure utility and to add utilities than in 1879 [Stigler, 1950, 317]. This feature appears from a passage in the first edition (p. 12) which he suppressed in the second: 'I confess that it seems to me difficult even to imagine how such estimations [of utility] and summations can be made with any approach to accuracy. Greatly though I admire the clear and precise notions of Bentham, I know not where his numerical data are to be found'. Eventually, Jevons arrived at his own conception of utility, as we shall see. Utility will become a precisely defined notion relating to goods and services.

A good's utility, Jevons affirmed (p. 51), is not an intrinsic quality of it, but merely the expression of the good's relation with mankind's pleasures and pain. It can be measured by a person's happiness; therefore it may differ between individuals. For every individual

Table 1. Contents by Chapters of Jevons's book. The titles of the chapters are given in italics.

Ch.: Page	**Titles and Description**
I: 3	*Introduction*. Mathematical character of the science; Confusion between mathematical and exact sciences; Capability of exact measurement; Measuring feeling and motives.
II: 33	*Theory of pleasure and pain*. Pleasure and pain as quantities.
III: 44	*Theory of utility*. Law of the variation of utility; Total utility and degree of utility; Variation of the final degree of utility; Distribution of commodity in different uses; Theory of dimensions of economic quantities; Distribution of a commodity in time.
IV: 79	*Theory of exchange*. Dimension of value; The law of indifference; The theory of exchange; Symbolic statement of the theory; Analogy to the theory of the lever; Complex cases of the theory; Competition in exchange; Failure of the equations of exchange; Negative and zero value; gain by exchange. Numerical determination of the laws of utility.
V: 162	*Theory of labour*. Quantitative notions of labour; Symbolic statement of the theory; Dimensions of labour; Relation of the theories of labour and exchange; Relations of economic quantities; Joint production.
VI: 198	*Theory of rent*. Symbolic statement of the theory; Illustrations.
VII: 212	*Theory of capital*. Quantitative notions concerning capital; Expression for amount of investment; Dimensions of capital, credit and debit; Effect of the duration of work; Illustrations of the investment of capital; General expression for the rate of interest; Dimension of interest; Advantage of capital to industry.
VIII: 254	*Concluding remarks*. Populations; Wages and profit; 'Obnoxious influence of authority'. [End 267.]

consumer, he distinguished *total utility* arising from a good and *degree of utility*, which has to do with the increase of utility caused by a (small) increase of the total quantity of the good and will be explained below.

Let the total quantity of food an individual consumes during a certain period be divided into ten increments indicated by roman numbers in Figure 1 (Fig. 3 on p. 55). If he were given only increments I and II he would possibly not starve, but obviously remain far from satiation. Increment III would therefore be very welcome and increase the individual's utility by a considerable quantity, measured by rectangle $pp'q'q$. The next increment's contribution is still considerable too, but somewhat less than $pp'q'q$; subsequent increments contribute lesser and lesser; the last one only slightly. Total utility is measured by the total surface of the ten rectangles. The law expressed by this figure is considered to be so generally felt by everyone that it apparently does not need further explanation or proof.

A mathematical problem is: what is measured along the y-axis of Figure 1? Jevons prudently gave no answer, but went on to the 'continuous case' (p. 58), with infinitesimally

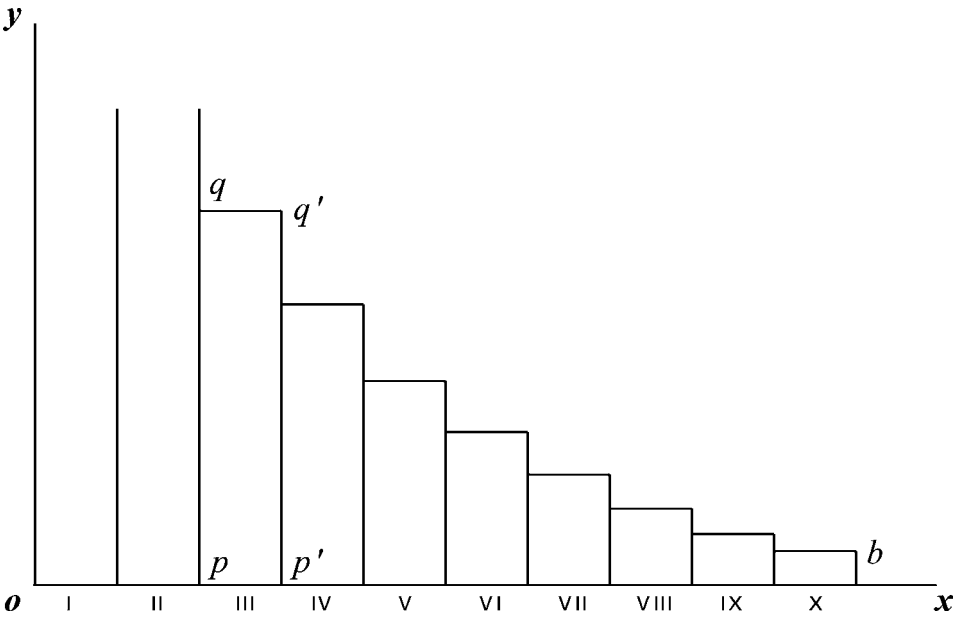


Figure 1. Law of Varying Utility.

small increments; see Figure 2 (Fig. 4 on p. 58). On the horizontal axis again the quantity of food is measured; on the vertical one he placed the 'degree of utility', and explained this as follows (pp. 58–59):

> When the quantity oa has been consumed, the degree of utility corresponds to the length of the line ab; for if we take a very little more food, aa', its utility will be the product of aa' and ab very nearly and more nearly the less is the magnitude of aa'. The degree of utility is thus properly measured by the height of a very narrow rectangle corresponding to a very small quantity of food, which theoretically ought to be infinitesimally small.

The length of ab is equal to the surface of $aa'b'b$ divided by the length of aa. Where this surface may be considered as a small increment of total utility u and the length of aa' as a small increment of the total quantity x of food consumed, he concludes that the degree of utility is represented by the fraction $\frac{\Delta u}{\Delta x}$. Because utility is considered to vary with 'perfect continuity', a small error is made in assuming it to be uniform over the whole increment. This error can be avoided if x is considered to be infinitesimally small. This leads to '*The degree of utility is*, in mathematical language, *the differential coefficient* $\frac{du}{dx}$ of u considered as a function of x, and will be itself another function of x' (pp. 60–61, italics in original). It is stated as a general law that the degree of utility is a decreasing function of x. The degree of utility of the last infinitesimally small unit consumed has a special name: 'final degree of utility'; it will play an important role below. Mathematically, if n is the total quantity consumed (see Figure 2), then the final degree of utility is measured by the length of nq. Note that these graphs may differ considerably from individual to individual.

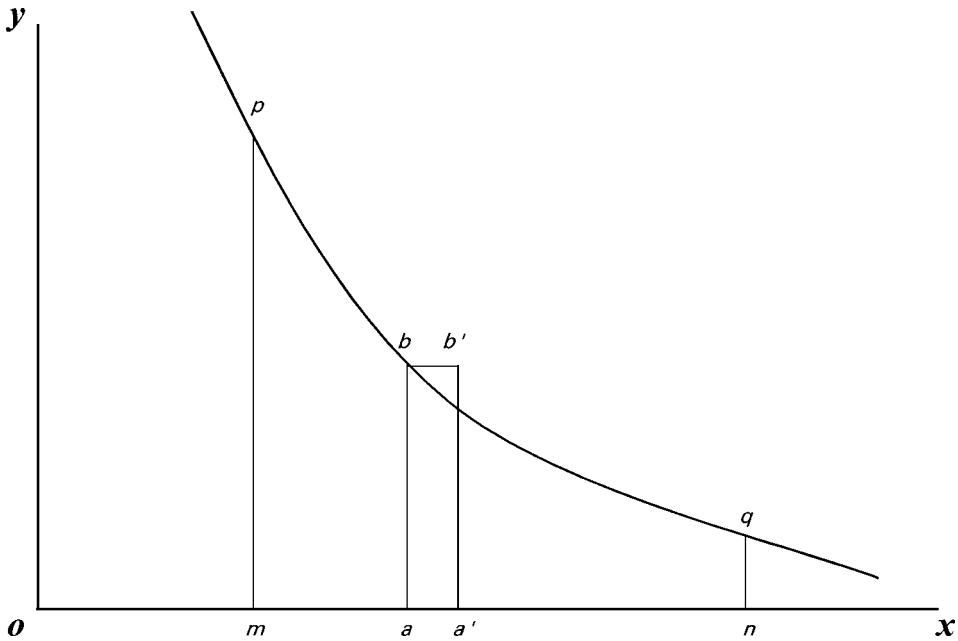


Figure 2. Degree of Utility.

We end this section by observing that an individual's total utility enjoyment of consuming the quantity n is measured in Figure 2 by the surface between the two co-ordinate axes, line nq and the graph of the degree of utility. One might also say that this total utility will be obtained by integrating the degree of utility, as a function of the quantity x, from 0 to n. In this set-up, total utility is therefore a derived notion; analysis does not start from it.

2 THE LAW OF EXCHANGE AND THE TRADING BODIES

According to Jevons, exchange takes place in markets. With this term is not meant some real place with real provisions to expose the goods for sale. It is rather a group of persons in business relations to transact intensively in one or more goods, for instance the world copper market. Jevons investigates traders' behaviour in what he calls 'perfect markets'. A market is 'theoretically perfect only when all traders have perfect knowledge of the conditions of supply and demand, and the consequent ratio of exchange' and 'there can only be one ratio of exchange of one uniform commodity at any moment' (p. 87). The stock exchanges, for instance, and other well-organised markets may be considered as good approximations of perfect markets. For obvious reasons, Jevons chose to analyse in first instance this type of market.

One of the most important elements in the analysis of exchange is value. There are several different notions of value, Jevons argued: 'value in use', expressed by total utility; 'urgency of desire for more', measured by the 'final degree of utility'; and as a third, pivotal notion 'purchasing power' or value in exchange of a good, i.e. the ratio of the

quantity of that good and the quantity of another good exchanged for it. The latter notion has everything to do with prices. It has cost economists a lot of time to realise firstly that the price of a good is not determined by its intrinsic properties, and secondly that the price of a good is always measured in another good. The latter fact means that the quantities, however great or small, of two goods exchanged must always have the same proportions in the same market. This brought Jevons to his 'Law of Indifference' (to quote his later name for it): 'The last increments in an act of exchange must be exchanged in the same ratio as the whole quantities exchanged' (p. 94). Let x and y be these quantities; then the law of indifference may be expressed as follows:

$$\frac{dy}{dx} = \frac{y}{x}. \tag{1}$$

Now we have all elements for Jevons's theory of exchange. We cannot give due to all the subtle details with which he dealt with in his book; unfortunately, it happens too often that economic theories are belittled on the basis of unkind simplifications which are then considered as representing the whole theory (see, for instance, Mirowski [1989]). So, if the reader wished to blame Jevons for being insipid, we should be blamed actually for misrepresenting his results.

With this proviso, we start with the simplest case. Let there be two individuals, A and B, who want to exchange corn and beef. A holds a quantity a of corn and no beef; B a quantity b of beef and no corn. If human nature's principles are correctly represented above, it is certain that exchanging a little corn for beef and beef for corn will increase both persons' total utility, since both will then give up a 'last increment' of one good for a 'first increment' of another (see Figure 1). How far will this exchange be continued? Jevons answers: 'The ratio of exchange of any two commodities will be inversely as the final degrees of utility of the quantities of commodity available for consumption after the exchange is effected' (pp. 95–96). He goes at length into this main proposition of the theory of exchange. Just as the other pioneers do, he presents as central argument that when the exchangers are in this point of equilibrium further exchange of small quantities of beef and corn will bring about a decrease of total utility for both. There is a paradoxical element in Jevons's exposition in the sense that he considers the price of a good measured in the other one, the reciprocal of the exchange rate, as more or less known to the exchangers, while at the same time this ratio is determined by the situation of equilibrium. This becomes clearer in his symbolic, mathematical statement of the theory.

Individual A, Jevons stated, will not be satisfied unless the following equation holds good:

$$\phi_1(a - x) \cdot dx = \psi_1 y \cdot dy, \tag{2}$$

where ϕ_1 and ψ_1 denote individual A's final degrees of utility for corn and beef as functions of the quantity of these goods. (Jevons did not write '$\psi_1(y)$'.) Note that final degrees of utility are here functions of a single variable: the quantity of the concerning good itself. All pioneers made this simplification, which means that total utility is an additively separable function of the quantities of the goods consumed. (The idea of generalising this was already in the air (for example in [Edgeworth, 1881, 20 ff.], where, incidentally, total utility is chosen as the basic concept, from which marginal utility is derived.) The quantities

exchanged of corn and beef are, respectively, x and y. A's quantities consumed are then $a - x$ and y. Because of the law of indifference we have again

$$\frac{dy}{dx} = \frac{y}{x}. \tag{3}$$

This ratio can be interpreted as the price of beef in corn.

If equation (2) holds good, is it not beneficial to A (in terms of utility) to exchange more corn for beef in quantities of the same proportion as (3), the only proportion admissible in this 'market'? This follows from the fact that, paraphrasing Jevons's mathematics,

$$\phi_1(a - x - dx) \cdot dx > \psi_1(y + dy) \cdot dy, \tag{4}$$

that is, utility lost, the left-hand member, exceeds utility gained.

Similarly, with self-evident symbols, B will be satisfied when

$$\psi_2(b - y) \cdot dy = \phi_2 x \cdot dx. \tag{5}$$

Combining (2), (3) and (5) yields (p. 100):

$$\frac{\phi_1(a - x)}{\psi_1 y} = \frac{y}{x} = \frac{\phi_2 x}{\psi_2(b - y)}. \tag{6}$$

These two equations with two unknowns, x and y, formulate Jevons's basic result in the theory of exchange. They do not just form an analogy in mathematical form of assertions already discussed verbally. This would not be enough for economics to be a veritable science. As all other sciences it must reason by equations that have a real meaning, to reach the position of a systematic science. Jevons started making economics a science in this sense.

Of course, equations (6) cover a very restricted situation only. They must be generalised in at least three directions. Firstly, goods exchanged are not always infinitely divisible, as assumed above; for instance, milk and tables cannot be dealt with similarly. Secondly, more than two goods are often involved in an exchange; moreover there is a very special commodity, namely labour, whose exchange deserves special attention. Thirdly, normally the number of exchangers in a market is much more than two.

Here we shall restrict ourselves to the last point because Jevons tried to solve it in a highly original but controversial way. In his set-up parties A and B are not necessarily individuals in the normal sense; the term 'trading body' is used to refer to them. A trading body may denote both a single individual and a group of individuals, for instance all the inhabitants of a certain country or all entrepreneurs in a certain branch of industry. In order to arrive at a formula like (6) above for all types of trading bodies, Jevons refers to his 'Principle of fictitious means' [1874, vol. 1, 422 ff.]. According to this principle, scientific laws that are theoretically correct for individual units separately are 'practically valid' for aggregates as well. (But his principle is dubious: in physics, for instance, it is easily refuted by the fact that the behaviour of oxygen molecules individually in a certain space, in terms of velocity is quite different from the characteristics of these molecules' aggregate, oxygen gas in terms of pressure and temperature; Boyle's law, for instance, has no micro analogue.) In his *Theory*, 90, he stated it thus:

> Thus, our laws of Economy will be theoretically true in the case of individuals and practically true in the case of large aggregates; but the general principles are the same, whatever the extent of the trading body considered. We shall be justified then, in using the expression with the utmost generality.

A modern version of Jevons's trading body is the fiction of the 'representative agent', supposed to act in much the same way as individuals do. A representative consumer, for example, is assumed to consume an amount of every good equal to the average amount consumed in the entire economy. The micro, per capita data are then simply multiplied by the number of consumers, and the products are put directly into the aggregate equations. This econometric practice may have some limited uses, provided that its numerical consequences are understood [Van Daal and Merkies, 1985, ch. 7]. The problems with the idea of a representative agent are that it is basically tautological and that it suppresses many individual differences that are of the essence of economic life. The idea is unrealistic and results in unrealistic models.

3 CONCLUDING REMARKS

The mathematics of Jevons was of the level of the average English scientist of his time. He was more interested in practical applications of mathematics to science rather than in mathematics as such. In the 1870s and 1880s rigour *à la* Cauchy (§25) and Gauss did not yet worry most scientists, and it was quite normal to treat differentials and differences in the same way, as we have seen above. In this connexion the following citation from the preface to the second edition (1879) of the *Theory*: 'In short, I do not write for mathematicians, nor as a mathematician, but as an economist wishing to convince other economists that their science can only be satisfactorily treated on an explicitly mathematical basis' (pp. xiii–xiv).

We pose a question that is still haunting economic theory: Will a situation as depicted by a solution of equations (6), or a generalization of them really be reached when starting from some initial situation, and how will it be reached? Indeed, such a solution is an equilibrium in the sense that, once reached no trader can change his position for a better one while taking account of the prices as determined by ratios like y/x. It is, however, a moot question whether two individual persons always will end up in a situation described by (6). If A and B are not in the same position as to power and if there are no institutions that prevent the one from abusing the other, than the outcome might be different, as the reader can easily imagine. If however A and B are multi-person trading bodies then there must also be more institutions in the markets, to increase the probability of a fair outcome of exchange.

Finally, we mention that Jevons's economic work was not restricted to the *Theory of political economy*. In particular, in a work published two years after his death as [Jevons, 1884], *Investigations in currency and finance*, he founded a theory on business cycles on the solar cycles.

Three years after Jevons's book appeared, there was published in Paris *Principe d'une théorie mathématique de l'échange* by Léon Walras (1834–1910). In spite of the different presentation, Jevons recognised a confirmation of his own ideas on utility. Walras [1874] immediately acknowledged Jevons's priority, while the latter regretted not to have deduced

demand functions from his functions of final degree of utility. Some years later both had to acknowledge that the priority Walras acceded to Jevons had to be passed back further, namely to Hermann Heinrich Gossen (1810–1858), whose book [Gossen, 1854] went nearly unnoticed until Jevons and Walras did justice to it. All this made a fine story in the Introduction to the second edition of Jevons's *Theory* (pp. xxvi–xl).

In Britain F.Y. Edgeworth, in *Mathematical psychics* [1881], continued Jevons's investigations regarding the problem of bilateral monopoly, and generalised utility functions. In his private publications, and later in his *Principles of economics* (1890), Alfred Marshall gave a different but esoteric orientation to the research started by Jevons: the curve of the price demanded points indirectly to marginal utility under the reservation that the marginal utility of money remains (approximately) constant [Marshall, 1920, Book IV, Chapter VI and Mathematical Appendix VI].

The expression 'neo-classical' was suggested by Thorstein Veblen in [1900] to designate essentially Marshall's theory. Later authors used it to indicate the tradition of marginalist work, in particular that of its founding fathers, Jevons, Menger and Walras.

BIBLIOGRAPHY

Bentham, J. 1789. *An introduction to the principles of morals and legislation*, London: Payne. [New ed. London: Wilson and Pickering, 1823.]

Black, R.D.C. 1962. 'Jevons, Bentham, and De Morgan', *Economica, n.s., 39*, 119–134. [Repr. in [Wood, 1988], vol. 1, 280–297.]

Edgeworth, F.Y. 1881. *Mathematical psychics*, London: Kegan Paul.

Ekelund, R.B. and Shieh, Y.-N. 1989. 'Jevons on utility, exchange, and demand theory: a reassessment', *The Manchester School, 62*, 17–33.

Gossen, Hermann Heinrich, 1854. *Entwickelung der Gesetze des menschlichen Verkehrs und der daraus fließenden Regeln für menschliches Handeln*, Braunschweig: Vieweg. [2nd ed. Berlin: Prager, 1889. 3rd ed. (intro. by F.A. von Hayek), Berlin: Prager, 1927. Reprints of 1st ed. Amsterdam: Liberac, 1967; Düsseldorf: Verlag Wirtschaft und Finanzen, 1987.]

Grattan-Guinness, I. 2002. ' "In some parts rather rough": a recently discovered manuscript version of William Stanley Jevons's "General mathematical theory of political economy" (1862)', *History of political economy, 34*, 685–726.

Howey, R.S. 1960. *The rise of the marginal utility school*, New York: Columbia University Press.

Inoue, T. and White, M.V. 1993. 'Bibliography of published works by W.S. Jevons', *Journal of the history of economic thought, 15*, 122–147.

Jevons, W.S. 1866. 'Brief account of a general mathematical theory of political economy', *Journal of the Statistical Society of London, 29*, 282–287. [Repr. in *Theory*, 4th ed. (1911), 303–314.]

Jevons, W.S. 1874. *Principles of science: a treatise on logic and scientific methods*, 2 vols., London: Macmillan.

Jevons, W.S. 1884. *Investigations in currency and finance* (ed. H.S. Foxwell), London: Macmillan.

Jevons, W.S. 1972–1981. *Papers and correspondence of William Stanley Jevons* (ed. R.D.C. Black and R. Konekamp), 7 vols., London: Macmillan.

Marshall, A. 1920. *Principles of economics*, 8th ed., London: Macmillan.

Mirowski, P. 1989. *More heat than light*, Cambridge: Cambridge University Press.

Peart, S. 1996. *The economics of W.S. Jevons*, London: Routledge.

Reid, G.C. 1972. 'Jevons's treatment of dimensionality in the theory of political economy: An essay in the history of mathematical economics', *The Manchester School, 40*, 85–98. [Repr. in [Wood, 1988], vol. 2, 70–83.]

Schabas, M. 1990. *A world ruled by number. William Stanley Jevons and the rise of mathematical economics*, Princeton: Princeton University Press.

Stigler, G.J. 1950. 'The development of utility theory', *Journal of political economy*, *58*, 307–327, 373–396. [Repr. in Stigler, *Essays in the history of economics*, Chicago: University of Chicago Press, 1965, 66–155.]

Van Daal, J. and Merkies, A.H.Q.M. 1985. *Aggregation in economic research: from individual to macro relations*, Dordrecht: Reidel (Theory and Decision Library, vol. 41).

Veblen, T.B. 1900. 'The preconceptions of economic science III', *Quarterly journal of economics*, *14*, 240–269.

Walras, L., 1874. 'Principe d'une théorie mathématique de l'échange', *Séances et travaux de l'Académie des sciences morales et politiques* (*Institut de France*), Extrait du Compte-rendu rédigé par M. Ch. Vergé, *Collection, new ser*. *101*, pt. 1, 97–116.

Wood, J.C. (ed.) 1988. *William Stanley Jevons: critical assessments*, 3 vols., London: Routledge.

CHAPTER 42

FELIX KLEIN'S ERLANGEN PROGRAM, 'COMPARATIVE CONSIDERATIONS OF RECENT GEOMETRICAL RESEARCHES' (1872)

Jeremy Gray

Klein's Erlangen Program was his review of contemporary methods in geometry. It became, some 20 years later, the work from which a new generation of mathematicians came to see how geometry was being done and to appreciate the importance of group theory in the study of geometry. The reason for this delay, and also for its subsequent and continuing impact, was the novelty with which Klein re-united the disparate fields of geometry through his emphasis on the role of groups of geometric transformations.

First publication. *Vergleichende Betrachtungen über neuere geometrische Forschungen. Programm zum Eintritt in die philosophische Facultät und den Senat der Friedrich-Alexanders-Universität zu Erlangen*, Erlangen: Deichert, 1872. 48 pages.

Further edition. With additional notes, *Mathematische Annalen*, *43* (1893), 63–100. [Repr. in Klein, *Gesammelte mathematische Abhandlungen*, vol. 1, Berlin: Springer, 1921, 460–497. Also in *Das Erlangen Program* (ed. H. Wussing), Leipzig: Geest & Portig, 1974 (*Ostwalds Klassiker der exakten Wissenschaften*, vol. 253).]

Italian translation by G. Fano, *Annali di matematica, (2) 17* (1890), 307–343.

French translation by H.E. Padé, *Annales de l'Ecole Normale Supérieure, (3) 8* (1891), 87–102, 173–199.

English translation by M.W. Haskell, *Bulletin of the New York Mathematical Society*, *2* (1893), 215–249.

Hungarian translation by Lajos Kopp, Budapest: 1897.

Russian translation by D.M. Sintsov, *Memoirs of the Physical–Mathematical Society of Kazan, (2) 5–6* (1895–1896), 16 + 28 pp.

Polish translation by S. Dickstein, *Prace matematyczno-fizycznych*, *6* (1905).

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

Related articles: Riemann on geometry (§39), Weber (§53).

1 ON THE BIOGRAPHY OF KLEIN

Christian Felix Klein was born in Düsseldorf, Germany, on 25 April 1849. His father was a local government officer; his mother, to whom Klein attributed the intellectual liveliness of his home, came from an industrial family in Aachen. Klein had a conventional German schooling in a Gymnasium, supplemented by some family friends and contacts who appreciated his quickness of mind, and he entered the University of Bonn at the age of sixteen and a half years, intending to study mathematics and natural science. However, the instruction there was at a very low level, until he had the good fortune to become Plücker's assistant. Julius Plücker (1801–1868) had followed a distinguished career as a mathematician, writing on the algebraic geometry of curves and resolving some of the central problems in projective geometry and duality. He then transferred his attention to experimental physics, and was one of the first to discover cathode rays, for which he was awarded the Royal Society of London's Copley Medal in 1866. He then switched back to pure mathematics and introduced the study of what is called 'line geometry'. Here the emphasis is on assigning coordinates to lines in space and defining configurations of lines by means of equations in these coordinates.

Klein was responsible for the demonstrations in Plücker's lectures on experimental physics, but he also assisted him with his researches in line geometry, which became the topic of his doctorate, which he finished in December 1868. Plücker had died, unexpectedly, in May that year, leaving his two-volume book on line geometry unfinished, and Klein was the ideal person to finish it. He therefore left Bonn and went to work with Rudolf Clebsch (1833–1872) in Göttingen, because Clebsch was undoubtedly the leading German geometer of his generation, and a stimulating figure around whom many others gravitated. A measure of how fast Klein had risen is given by the remarkable fact that even by the end of 1868 he had never heard a lecture on the integral calculus!

In Göttingen Klein learned more geometry than physics, and seeking to broaden his intellectual horizons travelled to study in Berlin, as almost all German mathematicians did at that time. There he could attend seminars in the largest and most powerful university for mathematics in the world, dominated by Leopold Kronecker, Ernst Kummer and Karl Weierstrass. The experience, however, was not congenial: the Berlin mathematicians placed more emphasis on mathematical analysis than Klein did, and paid more attention to rigorous arguments and special cases than Klein was ever to do. What redeemed the trip was his meeting the Norwegian mathematician Sophus Lie (1842–1899). Lie was a few years older than Klein, and like him lacking in the mathematical sophistication expected in Berlin. But they were united in the appreciation of geometry, in particular projective geometry and line geometry, and increasingly in the belief that they had new and valuable things to say in mathematics.

Klein and Lie began to do original research together, and together travelled to Paris to learn from the mathematicians there, especially the young geometer Gaston Darboux, and the group theorist Camille Jordan. Klein's trip was cut short by the outbreak of the Franco-Prussian war, and he returned home to serve in the German Army, acquire typhoid and be

invalided out, and make the acquaintance of one Friedrich Althoff, who was later to be the Minister of Education for Prussia and a considerable help to Klein in his academic career.

By 1871, Klein was back in Göttingen, where he concluded his *Habilitation* that January. He was torn between mathematics and physics, but Clebsch's influence was decisive because it was on his recommendation that Klein was appointed as a full Professor of Mathematics in Erlangen in autumn 1872, at the remarkably young age of 23. As was the custom, Klein had to present an Inaugural Address. This address was on the teaching activity he planned and the nature and purpose of an education in mathematics at all levels from school to university. It is commonly confused with the Erlangen Program, but that was not the Address [Rowe, 1983]. Rather, the Erlangen Program was a pamphlet printed by Deichert in Erlangen and distributed to those who came to the Inauguration. A few copies were doubtless distributed to friends and colleagues abroad, and to some libraries, because that was customary at the time; but the informal nature of the publication partially accounts for the negligible response to the Erlangen Program in 1872.

2 THE ERLANGEN PROGRAM

Klein chose the title to his essay carefully: he intended first to review, and then to compare, a number of recent researches in different areas of geometry. He claimed no novelty for the way he treated specific topics; what was original was the unified viewpoint he offered and its suggestions for the direction of future work. This viewpoint centred on the group-theoretic classification of the different geometrical methods then in use, and this emphasis owes a lot to Lie's influence. The Program was written while Klein was in daily contact with Lie, and reviewing its origins as he did when it was reprinted in his Collected Works he wrote that Lie was very much persuaded of the merits of the idea [1921, 411].

It is easiest to understand this idea in its paradigm example: the way non-Euclidean geometry appears as a sub-geometry of projective geometry. This was only done in an Appendix to the Erlangen Program, but it was developed at length in two papers published in 1871 and 1873 in the journal that Clebsch and others had recently founded, the *Mathematische Annalen*. Klein had learned of non-Euclidean geometry from Otto Stolz when in Berlin, but initially had had a hard time understanding it, and then in persuading Weierstrass of the value of his new point of view.

Inspired by a serendipitous reading of a paper by the English mathematician Arthur Cayley, who had had a glimmer of the same idea, Klein argued as follows. The model of non-Euclidean geometry developed by Eugenio Beltrami draws the entire non-Euclidean plane inside a circle, and it draws straight lines in non-Euclidean geometry as straight lines inside the circle (§39.5). This suggested to Klein that the allowable transformations of figures in non-Euclidean geometry ought to be those which are projective transformations mapping the circle to itself, because they will automatically map each straight line to a straight line. Now, a non-Euclidean transformation is one that maps a line segment in non-Euclidean geometry to another line segment of equal non-Euclidean length. Projective geometry, on the other hand, can map any two points to any two points; the most important property of projective transformations is that they map four points on a line to four points on a line if and only if the four points have the same cross-ratio. For Klein's idea to

work, he had to find a way of expressing non-Euclidean distance, which only involves two points, in terms of the four point projective invariant, cross-ratio. He did this by observing that two points inside the circle define a non-Euclidean straight line that meets the circle in two more points, thus giving him four points. It remained for Klein to define non-Euclidean distance between the original two points in terms of the cross-ratio of the four points, and this he did by a straightforward technical argument.

The upshot of all of this was that Klein had two geometries, projective geometry and non-Euclidean geometry, and each geometry had a family of allowable transformations. In fact, each of these families is what is technically called a group of transformations. Moreover, the space of non-Euclidean geometry is a subspace of projective geometry, consisting of the points inside a fixed but otherwise arbitrary conic (for convenience, a circle) and the transformation group of non-Euclidean geometry is a subgroup (in the technical sense of the term, to be defined below) of the transformation group of projective geometry. So, simply but accurately, non-Euclidean geometry could be called a sub-geometry of projective geometry.

Klein began the Erlangen Program by alluding to the recently discovered relation between the metrical and projective properties of figures, but in the more natural case of Euclidean geometry. Then came the geometry, as he called it, of reciprocal radii vectores, today called inversive geometry, and then birational geometry, which will be defined below.

Having introduced the main geometries, Klein then introduced the group concept. This was not widely known in 1872. His presentation, as he himself noted when the work was reprinted in the 1890s, was less than perfect: the only property of a group that he insisted upon was closure. Perhaps the simplest example of a group is the integers with the operation of addition. More generally, and in modern terms, a set of objects forms a group if:

- closure holds: any two objects can be combined to form a third which also belongs to the group (in symbols $A + B = C$);
- there is an identity object, often denoted e, with the property that the combination of any object with the identity returns that object ($A + e = A$);
- every object, say A, in the group has an inverse, usually denoted $-A$, which is an object such that the combination of an object and its inverse is the identity ($A + -A = e$); and
- the so-called associative law holds: $A + (B + C) = (A + B) + C$.

A subgroup is a subset of a group that is also a group. The integers form a group because the sum of two integers is a group, the integer 0 is the identity element, the inverse of the integer n is the integer $-n$, and the associative law holds: $k + (m + n) = (k + m) + n$. In Euclidean geometry, the group of all distance-preserving transformations consists of the familiar rotations, translations, reflections and their composites.

Klein may have skimped on the definition of a group because he did not do something later generations were to do, namely he did not distinguish between an abstract group and a transformation group. For groups whose elements are transformations it is, for example, automatic that the associative law holds, and it is usually easy to see if a set of transformations contains an identity element and an inverse for every element. But these imperfections

should not obscure the magnitude of the point Klein was insisting upon. No-one had disputed for centuries that geometry, any geometry, involved the use of transformations, for example to replace a figure with an equivalent but simpler one, or to choose more convenient co-ordinate axes. Klein shifted mathematician's attention from the figures to the transformations, and argued that geometry was about groups as well as the properties of shapes.

In Sections 1–3 of the Erlangen Program, Klein discussed the use of the group concept, in particular the idea of one group being a subgroup of another. This enabled him to fix a space and vary the group, either to introduce a new geometry or to recognise a known one in an unexpected setting. This enabled him to distinguish usefully between the invariants of one group and another, and therefore between their geometrical properties, as the paradigm example of non-Euclidean geometry and projective geometry demonstrates. But Klein also pointed out that spaces could not only be of any dimension, but the points of one could be objects in another: one might study the space of all lines in a projective plane, or the space of all conics. This transition took one into what his contemporaries called higher geometry, which Klein had learned from Plücker and then greatly extended. The example of introducing imaginary elements (those described by complex coordinates) was mentioned explicitly, and motivated by the desire to bring geometry into line with algebra.

In Section 4, Klein discussed a way in which two seemingly distinct geometries can be shown to be the same. This is the method of transfer, as he called it, which applies when there is an invertible transformation, t say, between two spaces A and A', with transformation groups B and B' respectively. The map between B and B' that sends an element b of the group B to the element tbt^{-1} (where t^{-1} is the inverse of t) in the group B' transfers the group acting on the space A to the space A'. If the action of the transferred group is identical with that of the group B'—for which the technical term is isomorphic—then the geometries could reasonably be said to be equivalent. Similarly, such a transfer can introduce a geometry onto a space by allowing one to transfer the action of a group on a space to the action on a new space.

Klein gave the example where the space A is a projective line and the group B is that of the rational transformations. The points of the projective line have coordinates of the form $[x, y]$ where $[x, y]$ and $[x', y']$ represent the same point if and only if there is a non-zero rational number k, say, such that $x = kx'$ and $y = ky'$, and the expression $[0, 0]$ does not represent any point of the projective line. A projective transformation is of the form $[x, y] \to [ax + by, cx + dy]$, with $ad - bc \neq 0$. This group also acts on binary forms, which are expressions of the form $\alpha x^2 + \beta xy + \gamma y^2$, so the geometry of the projective line is the same as the study of binary forms. Now, the line is obviously a geometrical object, but the space of binary forms is not, and this was, for Klein one of the advantages of the Erlangen Program: it enabled one to introduce the resources of geometry into a branch of mathematics previously regarded as purely algebraic.

Klein then discussed the transfer of the projective geometry of the plane onto a quadric by stereographic projection. As a good 19th-century geometer, he did not care if the spaces were real or complex: his example is easier for us to understand if the space is real and the quadric is a hyperboloid of one sheet. The transfer map picks as a centre of the projection a point on the hyperboloid, and maps the two straight lines of the hyperboloid through that point onto points at infinity in the image projective plane. Any other point of the

hyperboloid is mapped onto the projective plane by the line joining it to the centre of projection. Thus elementary plane projective geometry and the geometry of a quadric with a marked point are the same. In Section 5, the same process of transfer was illustrated in more elaborate spaces, and Klein carefully pointed out now that it is not enough that the spaces have the same dimensions; the group actions must also agree.

In Section 6 Klein looked at inversive geometry. In the plane (or, respectively, space) this is the geometry of circles (respectively, spheres) in the plane, and the allowed transformations are those mapping circles to circles (respectively, spheres to spheres), which can also be characterised algebraically. The inversive geometry of the plane, he pointed out, is identical to projective geometry on a quadric, while the inversive geometry of space coincides with the projective geometry on a quadric in projective five-dimensional space.

Section 7 of the Erlangen Program was devoted to mention of Lie's sphere geometry. This was an ingenious way of studying all spheres in three-dimensional space, by showing that the geometry was equivalent to the geometry of all lines in three dimensional space. Lie's line-sphere transformation turned out to have implications for differential geometry that enabled both Klein and Lie to have interesting new results in the early 1870s. Section 8 was yet more ambitious; Klein indicated that one might hope for a birational geometry of space, where the transformations are quotients of polynomials; and for topology (which he called 'analysis situs'), where the transformations are invertible continuous maps. He even held out the prospect of a geometry of all invertible differentiable maps, which would map (in modern language) the tangent space of a surface to itself. This idea led him in Section 9 to the study of contact transformations, which was to become a major theme of Lie's work in future years.

The 10th and final Section was devoted to some a few short remarks about other possibilities. One was the study of manifolds of constant curvature; another was derived from a comparison with the Galois theory of equations, leading to the suggestion that it would be fruitful to pass down a chain of subgroups.

The Erlangen Program ended with a series of seven notes of varying length and significance. Note 5 was on the 'so-called non-Euclidean geometry', as Klein cautiously continued to call it in order to avoid debates with non-mathematicians. Note 6 discussed Klein's work on line geometry, and clarified the tricky point that the projective geometry of space with respect to a fixed quadric does not impose a geometry on the quadric itself. The final Note hinted at a theme of growing importance to Klein: the geometrical interpretation of the invariants associated to a binary form.

The central message of the Erlangen Program is that every geometry is to be thought of as a space and a group of transformations acting on that space that preserve the essential features of that geometry. These might be metrical, as they are in Euclidean geometry or non-Euclidean geometry, the property of being a circle (inversive geometry in the plane) or of being a straight line (projective geometry). Two implications follow from this insight. One is that all the different geometries known at that date can be thought of as special cases in a hierarchy of geometries, thus re-unifying geometry, a subject that Klein thought had diversified too much. The other is that, by the principle of transfer, seemingly different geometries can be seen to be essentially the same, and a space can perhaps be given a geometry where previously it had none.

3 THE RECEPTION OF THE ERLANGEN PROGRAM

For an interesting pair of contrasting views of the question of the influence of the Erlangen Program see [Birkhoff and Bennett, 1988] and [Hawkins, 2000, 34–42]; and works cited there give even more strongly contrasting views. It seems that the Erlangen Program met with a slow reception until the 1890s, by which time Klein's status as a major mathematician at the University of Göttingen had a great deal to do with its successful re-launch. By that time too a number of mathematicians had done considerable work broadly in the spirit of the programme, although the extent to which they were influenced by the programme, or were even aware of it, is not at all clear.

Klein's remarkably active life may be divided into a number of phases. Almost from his arrival in Bonn in 1865 he was devoted to research. This phase ended with his collapse in 1882 from acute nervous exhaustion brought on by his attempt to match the achievements of the French mathematician Henri Poincaré (1854–1912) on a topic of mutual interest (complex function theory, group theory, and non-Euclidean geometry). By 1886 Klein had recovered his health, but his research became less original and more didactic. He had worked his way back into the subject with his famous book on the icosahedron [Klein, 1884], which is a recapitulation and reworking of earlier ideas. Thereafter he liked to work as a collaborator and a supervisor, with a gifted and usually younger colleague working through the detailed, technical aspects that had never been Klein's forte. In 1892, when Klein became a senior professor at Göttingen, he began to move slowly away from research and to pursue the organisational side of mathematics. Klein now became the most influential mathematician 'behind the scenes' and was extraordinarily successful in establishing Göttingen as the pre-eminent University for mathematics in the world. He made a number of inspired hirings, David Hilbert among them, organised a series of important investigations into the teaching of mathematics, and oversaw the production of a 23-volume *Encyklopädie der mathematischen Wissenschaften* ('Encyclopaedia of mathematical knowledge'), which was also partly translated into French, including wholesale up-datings and re-writings as appropriate. Klein's position, and the large number of students he had passing through Göttingen, undoubtedly contributed to the new reputation of the Erlangen Program.

It was at this stage, motivated by a desire to make his early work, including the Erlangen Program, better known and to remind the mathematical community of his old association with Lie, that Klein sought to republish their early work together. Since 1872 Lie had gone on to build up a vast theory of groups of continuous transformations of various kinds; but however much it owed to the early experiences with Klein, and however much Klein may have assisted Lie in achieving a major professorship at Leipzig University in 1886, it is doubtful if the Erlangen Program had guided Lie's thoughts. Lie was far too powerful and original a mathematician for that. By the 1890s he was also suffering from the early stages of an illness that alarmed many who knew him, for he became more and more prone to fits of anger.

Whatever the reason, Klein's attempts to involve Lie in the re-edition of the Erlangen Program back-fired terribly. Lie took the occasion of the publication of the third and final volume of his *Theorie der Transformationsgruppen* to say in the preface [Lie, 1893, 17]:

> I am not a student of Klein's nor is the reverse the case, even if it comes closer to the truth. [. . .] I value Klein's talent highly and will never forget that part he had in accompanying my scientific efforts from the beginning. I believe, however, that he does not always distinguish sufficiently between induction and proof, between the introduction of a concept and its utilization.

Remarks like this shocked the German mathematical community, and embittered the last few years of the relationship between Klein and Lie, who died in 1899. Nevertheless Klein was generous enough in 1897 to persuade the Physico-Mathematical Society of Kazan in Russia (who had just published a translation of Klein's essay) to award its first Lobachevsky prize to Sophus Lie for that volume of his *Theorie der Transformationsgruppen*.

Another mathematician who had done more on the connections between groups and geometry than Klein ever managed was Poincaré. He had almost certainly come to the idea that groups, and groups of transformations in particular, were fundamental mathematical objects independently of Klein. By 1880 he was clear that to speak of a geometry is to speak of a group, and he studied non-Euclidean geometry in this spirit (using a more metrical, less projective version of it, which is one reason any influence of the Erlangen Program is unlikely). As Poincaré put it: 'In fact, what is a geometry? It is the study of the group of operations to which one can subject a body without deforming it'. Poincaré went on to pioneer the introduction of group theoretic and geometric methods into complex function theory; he introduced vast classes of new functions into mathematics, some named 'Fuchsian' after the German mathematician Lazarus Fuchs, whose work had been a starting point for Poincaré's own, and some Poincaré called 'Kleinian' largely because Klein had objected to the name 'Fuchsian'.

But if the achievements of Lie and Poincaré had not been much inspired by the Erlangen Program, they were nonetheless powerful arguments for the value of bringing together group theory and geometry. What they did not do was exemplify the merits of the idea of a hierarchy of geometries, or of transferring a geometry from one space to another. Klein could justifiably claim some prescience in the Program, provided he did not insist too much on the details. It was in this attenuated, more general sense that the Erlangen Program was to exert a new influence.

Klein now oversaw the translation of the Erlangen Program into several languages (see the publication history at the head of this article). The young Italian Gino Fano, who had visited Göttingen, was persuaded to translate it into Italian (compare [Fano, 1907]). One of Klein's American students, M.W. Haskell, translated the Program into English.

Three stages in the shifting fortunes of geometry will help us understand the later life of the Erlangen Program. In 1899 David Hilbert proposed the first new departure in ways of thinking about geometry that went significantly beyond the Kleinian view (§55). In his *Grundlagen der Geometrie* he argued for an axiomatic presentation of geometry and for a comparison between geometries treated as systems of axioms. This left no prominent position for group theory. However, when Einstein came up with his special theory of relativity in 1905, Klein was very happy to see that it fitted directly into his Erlangen Program. The feeling that the Erlangen Program somehow captured the 'meaningful' geometries was installed in the modern theory of differential geometry, which grew up in the years following

Einstein's introduction of his general theory of relativity, by the work of Elie Cartan, Hermann Weyl, and others (§63). They showed that it made sense to define various forms of geometric structure on a manifold, corresponding to measurements taken in different but overlapping regions of the manifold, and that when this is done the infinitesimal structures that result are almost exactly those described in the Erlangen Program. The limited choice of groups that can arise derives from the work of Sophus Lie, but the association of a group with a geometric property is Klein's original vision greatly generalised and transferred to a new setting. The result is a potent vision of mathematics that embraces modern teaching and modern research, and continues to reflect well upon the Erlangen Program.

BIBLIOGRAPHY

Birkhoff, G. and Bennett, M.K. 1988. 'Felix Klein and his "Erlangen Program" ', in W. Aspray and P. Kitcher (eds.), *History and philosophy of modern mathematics*, Minneapolis: University of Minnesota Press (Minnesota Studies in the Philosophy of Science, vol. 11), 145–176.

Fano, G. 1907. 'Kontinuierliche geometrische Gruppen. Die Gruppentheorie als geometrischen Einteilungsprinzip', in *Encyklopädie der mathematischen Wissenschaften*, pt. 4, sec. 1, Leipzig: Teubner, 289–388 (article IIIAB4b).

Hawkins, T. 2000. *Emergence of the theory of Lie groups*, New York: Springer.

Hilbert, D. 1899. *Grundlagen der Geometrie*, 1st ed., Leipzig: Teubner. [Many rev. eds. from 1903; see §55.]

Klein, C.F. 1884. *Vorlesungen über das Ikosaeder und die Auflösung der Gleichungen vom fünften Grade*, Leipzig: Teubner. [Repr. (ed. P. Slodowy), Birkhäuser: Basel, 1993. English trans.: *Lectures on the icosahedron* (trans. G.G. Morrice), 1st ed., London: Kegan Paul, 1888. 2nd ed. 1913; repr. New York: Dover, 1956.]

Klein, C.F. 1921, 1922, 1923. *Gesammelte mathematische Abhandlungen*, 3 vols., Berlin: Springer.

Laptev, B.L. and Rozenfel'd, B.A. 1996. 'Geometry', in A.N. Kolmogorov and A.P. Yushkevich (eds.), in *Mathematics of the 19th century. Geometry, analytic function theory*, Basel: Birkhäuser, 1–117. [Russian original 1981.]

Lie, S. 1893. *Theorie der Transformationsgruppen*, vol. 3, Leipzig: Teubner.

Poincaré, H. 1997. *Three supplementary essays on the discovery of Fuchsian functions* (ed. and intro. J.J. Gray and S.A. Walter), Berlin: Akademie Verlag; Paris: Blanchard.

Rowe, D.E. 1983. 'A forgotten chapter in the history of Klein's Erlanger Programm', *Historia mathematica, 10*, 448–454.

CHAPTER 43

RICHARD DEDEKIND, *STETIGKEIT UND IRRATIONALE ZAHLEN* (1872)

Roger Cooke

This short work marks a significant epoch in the movement known as the arithmetization of analysis, that is, the replacement of intuitive geometric notions by concepts described in precise words.

First publication. Braunschweig, F. Vieweg & Sohn, 1872. 31 pages. [Repr. 1892, 1905 and 1912, and then posthumously.]

Reprint. *Gesammelte mathematische Werke* (ed. R. Fricke and others), Braunschweig: Vieweg, 1930 (repr. New York: Chelsea, 1969), 315–334.

English translation. *Essays on the theory of numbers* (trans. W.W. Beman), Chicago: The Open Court Publishing Company, 1901 (repr. New York: Dover, 1963), 1–27. [Repr. in F.W. Ewald (ed.), *From Kant to Hilbert. A source book in the foundations of mathematics*, 2 vols., New York and Oxford: Clarendon Press, 1996, 766–779.]

Russian translation. *Neprerivnost' i irratsionalnie chisla* (trans. S. Shatunovskii), Odessa: 1908.

Italian translation. *Essenza e significanza dei numeri, Continuità e numeri irrazionali* (trans. and ed. Oscar Zariski), Rome: Alberto Stock, 1926, 119–153.

Related articles: Riemann on trigonometric series (§38), Cantor (§46), Dedekind on the integers (§47).

1 INTRODUCTION: THE PROBLEM OF INCOMMENSURABLES

This essay has been frequently reproduced, often in conjunction with his *Was sind und was sollen die Zahlen?* (translated as *The Nature and meaning of numbers*), which is described in §47. Dedekind wrote this essay to clarify a foundational problem that had lain beneath the surface of analysis since the time of Descartes. The roots of the problem go back to the

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

earliest mathematics that we now recognize as formally deductive, to the problem of incommensurables, first raised in the fourth century BCE. The Pythagoreans had discovered that some ratios of lines, such as the diagonal and side of a square or regular pentagon, could not be expressed as ratios of integers. In order to save the theory of geometric proportion, a definition ascribed to Eudoxus of Cnidos was adopted by Euclid, based on the idea that a given multiple of one geometric object (length, area, volume, or weight) must be less than, equal to, or greater than a given multiple of another object of the same type. One could then construct a theory of proportionality by defining the proportion $a : b :: c : d$ to mean that, for any positive integers m and n, whatever relation ($ma < nb$, $ma = nb$, $ma > nb$) holds between a and b must also hold between c and d. In this way it was possible, in effect, to assert that two geometric ratios are equal without having to say what a geometric ratio *is*. Nowadays the missing definition could be supplied by defining a ratio to be an equivalence class of pairs of geometric objects. Obviously such a definition would not have occurred to the mathematicians of Euclid's time, and they were content to say that two quantities *had a ratio* without saying what the word *ratio* meant in isolation.

The solution given by Eudoxus sufficed for the purposes of geometry, and was soon simplified sufficiently to be used easily in proofs; but it introduced a separation between numbers and space that seemed insurmountable. The continuous and the discrete seemed to be irreconcilable. A distinction was made between *number* and (continuous) *magnitude*. For good reasons, Euclid gave separate discussions of the theory of arithmetic proportion in Books VII–IX and geometric proportion in Books V–VI.

The rise of algebra in the Islamic world a millennium ago and in European mathematics during the 16th century brought this problem once more to the fore. In general one expected the solution of an equation to have a numerical representation, but in many cases the algebraic operations to be performed could be represented only geometrically. Omar Khayyam, for example, had shown how to solve many cubic equations by intersecting conic sections, but never fulfilled a promise he had made to present numerical solutions of these equations. In fact, attempts to do so merely lead back to the original cubic equation. It appeared that the unknowns and constants in equations had to be interpreted as *magnitudes* rather than *numbers*, in order for a solution to be found. The interpretation of a magnitude as a line, and the product of two magnitudes as an area forced Omar Khayyam to adhere to a dimensionally homogeneous notation, in which each term contained the same number of factors. Each single factor represented a length, the product of two factors an area, the product of three factors a volume, and he said explicitly that the square of a square was meaningless. Pierre Fermat, who invented analytic geometry independently of Réné Descartes, always observed these rules of dimensionality in his equations. The problem was that geometric magnitudes had never been systematized in terms of arithmetic rules, since they had never been thought of as numbers.

It was finally Descartes who removed the need for dimensional homogeneity in the terms of an equation by showing how to interpret the product of two lengths as a length. He chose a fixed length I as a unit and defined the product ab as the length that satisfies the proportion $I : a :: b : ab$. A modern mathematician would say that Descartes had made equivalence classes of directed line segments into the field of real numbers. However, such an interpretation would have been nonsense to Descartes. He was concerned with getting a

geometric representation of algebraic operations, and would not have thought of applying the term *number* to the result.

The success of Descartes's analytic geometry and the calculus whose invention followed close on its heels led to the subject now called 'analysis'. This was an amalgam of algebra and geometry, but included also numerical interpretations of its results. In particular, one of the great advantages of analysis was the existence of Taylor series to make possible the approximate computation of transcendental functions such as the exponential and trigonometric functions. But the route to these series went through algebra and calculus and was therefore not purely numerical. Moreover, the infinitesimal arguments used were controversial for some time. Theorems such as the mean-value theorem of calculus depended on a notion of continuity to guarantee that a curve containing points on both sides of a line must intersect the line. But, as the Pythagoreans had shown, the numerical version of this theorem was false: the point of intersection might very well not correspond to any number. It was incorrect to call the intersection an irrational *number*, since there was no articulated theory of irrational magnitudes that allowed them to be added or multiplied like numbers. Algebraic rules such as $\sqrt{ab} = \sqrt{a}\sqrt{b}$ applied to magnitudes by *fiat*, but were difficult to prove, even geometrically; and no one had produced any corresponding arithmetic rules, or even a non-tautological definition of the square root of a non-square integer. Any such definition first of all begged the question of the *existence* of the object defined; and if existence is granted by appeal to geometry, the rules for treating lengths as numbers still needed to be formulated and proved correct.

Such was the situation that confronted Dedekind when he began teaching in Zürich in 1858. Irrational magnitudes could be compared with (rational) numbers, and thus must in some sense behave like numbers, but no one had given a definition of what we now call a real number and shown how such numbers were to be added and multiplied independently of the geometric interpretation Descartes had given. Dedekind was interested in the problem because of the pedagogical implications of these gaps in the literature, and also because of his abiding interest in foundational issues.

2 THE AUTHOR

Richard Dedekind was born on 6 October 1831 in Braunschweig, the son of a professor at the Collegium Carolinum; his mother's father was also a professor there. He never married, and he lived with his unmarried sister for most of his life. He attended the Collegium Carolinum from 1848 until 1850, when he entered the University of Göttingen. There he studied with Gauss, receiving the doctoral degree as the last student of Gauss in 1852. Since his education had proceeded at a rapid pace, Dedekind naturally had some gaps to fill. Over the next few years he filled these gaps, receiving his *Habilitation* alongside Bernhard Riemann in 1854. He then began teaching at Göttingen. Gauss died the following year and was replaced by J.P.G. Lejeune-Dirichlet, who became a good mentor for both Riemann and Dedekind. Riemann, however, was the senior partner of the two younger mathematicians, and Dedekind earnestly sought to learn about elliptic functions from him.

In 1858 Dedekind was chosen for a chair at the Zürich Polytechnikum and began teaching there in the fall of that year. It was to be his lot to be involved in editing the collected

papers of his three great mentors Gauss, Dirichlet, and Riemann, and some of his ideas were no doubt inspired by this work. He introduced the concept of an ideal in a ring in his 1871 edition of Dirichlet's lectures on number theory. In his history of 19th-century mathematics, Felix Klein objected that the word 'ideal' was a misnomer, since the kinds of principal ideals that Dedekind was interested in were quite 'real', having a concrete representation as the set of all multiples of a fixed element [Klein, 1926, 322].

Dedekind was also interested in foundational and philosophical questions. He made the acquaintance of the young Georg Cantor in 1874, and was very much in sympathy with the freedom of definition Cantor was trying to introduce into mathematics, so much so that he became both a founding contributor and a champion of set theory ([Ferreirós, 1999], and §46). His principal works are his edition of Dirichlet's *Vorlesungen über Zahlentheorie* first in 1863: §37) and a *Theorie der ganzen algebraischen Zahlen* of 1879, which contained most of his original contributions to modern algebraic number theory [Dedekind, 1964].

Dedekind became a member of several academies of sciences, including those at Paris and at Göttingen. He died in Braunschweig on 2 February 1916.

3 DEDEKIND'S VIEW OF THE PROBLEM OF CONTINUITY

This problem of continuity persisted even after A.L. Cauchy and Bernard Bolzano had removed some of the difficulties connected with the use of infinitesimal arguments (compare §25). It was a problem of interpretation. If the final output of a problem was to be a computable quantity, then the algebraic symbols from which the quantity was constructed required some numerical interpretation. But it was manifest that wherever continuity arguments were invoked an appeal was being made to geometric intuition. Such was the problem that Dedekind faced, starting in 1858. As he said in the introduction to his pamphlet:

> As professor in the Polytechnic School in Zürich I found myself for the first time obliged to lecture upon the elements of the differential calculus and felt more keenly than ever before the lack of a really scientific foundation for arithmetic. In discussing the notion of the approach of a variable magnitude to a fixed limiting value, and especially in proving the theorem that every magnitude which grows continually, but not beyond all limits, must certainly approach a limiting value, I had recourse to geometric evidences. [. . .] The statement is so frequently made that the differential calculus deals with continuous magnitude, and yet an explanation of this continuity is nowhere given. [. . .] It then only remained to discover its true origin in the elements of arithmetic and thus at the same time to secure a real definition of the essence of continuity. I succeeded Nov. 24, 1858, and a few days afterward I communicated the results of my meditations to my dear friend Durège with whom I had a long and lively discussion [. . .].

Dedekind did not regard his discovery as anything but a commonsense clarification that anybody could have made (see his 1876 letter to Rudolf Lipschitz, quoted below). For 14 years he made no attempt to publish the result, and then published it merely as a tribute to

his father on the occasion of 50 years in service. Dedekind took for granted the validity of the geometric intuition on which much of calculus had been based. His meditations were concerned with formulating the geometric principle involved in precise language. As he discovered when he succeeded, the principle was in fact a tacit assumption of geometry made by all geometers since Euclid, and no one had suspected that this assumption was being made over a period of more than two thousand years. In putting a better foundation under the calculus, Dedekind also helped to shore up the foundations of geometry.

4 DEDEKIND'S SOLUTION OF THE PROBLEM

In the preface to his essay, Dedekind paved the way for the introduction of a new kind of number. The ground for such a creative act, he argued, had been broken previously as mathematicians invented negative, fractional, and imaginary numbers. In the first section of the essay (properties of the rational numbers) Dedekind mentioned in passing some work he himself had done, which is now recognized as one of the foundations of modern algebraic number theory, inventing the abstract object known as a *number field* ('*Zahlkörper*'), which is a 'system' R such that the four arithmetic operations can always be performed on any two 'individuals' of R, the only exception being division by zero. The use of the words 'system' and 'individuals' is significant, as mathematicians would now much more naturally use the words 'set' and 'elements' or their synonyms.

Dedekind's work was no small contribution to abstract set theory, which had not yet crystallized in his work and that of Eduard Heine and Cantor (compare §46). He remarked in his preface that he had just received Cantor's paper generalizing Riemann's uniqueness theorem for trigonometric series representations [Cantor, 1872] and was pleased to note that Cantor was using an axiom of continuity compatible with his own. As a matter of fact, Cantor defined a real number, which he called a 'numerical magnitude' ('*Zahlgrösse*'), as what is now called a Cauchy sequence of rational numbers, identifying the real number defined by two such sequences if the difference between the sequences tended to zero. It is difficult to see similarities in the two approaches, except that both involved inserting newly created numbers among the rational numbers. Cantor, however, agreed that the two approaches were compatible, saying to Dedekind that 'only in the conceptual introduction of [real numbers] is there any difference. I am fully convinced that the essence of continuity consists of what you have presented'.

The first Section concludes with a discussion of the three order properties of rational numbers, in which Dedekind was to find the secret of defining the whole set of real numbers: 1) The transitivity of the 'greater than' relation; 2) infinite divisibility (between any two rational numbers there exist infinitely many rational numbers); and 3) the separation property, which says that each rational number a separates the whole class into two distinct sets of numbers, those that are smaller than a, and those that are larger. By the first property, each number of the first class is smaller than each number of the second class, and the number a itself could then be assigned to either class, as either the largest number in the first class or the smallest number in the second.

In the second Section, on the comparison of the rational numbers with the points of a straight line, Dedekind drew the formal analogy between these properties and the order

properties of a geometric line: 1) the relation 'right of' is transitive, that is, if p is right of q and q is right of r, then p is right of r; 2) for any two points on a line there are infinitely many points between them; and 3) each point p on a line divides the line into two classes, those to the left of p and those to the right, and the point p can be assigned to either class as the rightmost point of the first or the leftmost point of the second. Selecting an origin and a unit of length on the line (and a positive direction, although Dedekind did not mention this fact) would establish a real correspondence between these two sets of order properties, so that each rational number would correspond to a point on the line, and the two classes into which the number separates the set of rational numbers would correspond to the two classes into which the corresponding point separates the line.

In the third Section of the essay, on the continuity of the straight line, Dedekind reaped the harvest of these seemingly simple considerations, which, as he said (pp. 9–10),

> are so familiar and well known to all that many will regard their repetition quite superfluous. Still I regarded this recapitulation as necessary to prepare properly for the main question. For, the way in which the irrational numbers are usually introduced is based directly upon the conception of extensive magnitudes—which itself is nowhere carefully defined—and explains number as the result of measuring such a magnitude by another of the same kind. Instead of this I demand that arithmetic shall be developed out of itself.

In a footnote he pointed out that this appeal to measuring one magnitude by another of the same kind fails when applied to complex numbers. And, he said, in any case the idea could be made clear only *after* a theory of irrational numbers was constructed. By making these commonplace remarks explicit, Dedekind was at last able to exhibit the hidden assumption of geometry that had been used in both geometry and analysis for centuries. That property was the converse of the third property listed above (pp. 11–12):

> I find the essence of continuity [...] in the following principle: *If all points of the straight line fall into two classes such that every point of the first class lies to the left of every point of the second class, then there exists one and only one point which produces this division of all points into two classes.* [...] As already said I think I shall not err in assuming that everyone will at once grant the truth of this statement; the majority of my readers will be very much disappointed in learning that by this commonplace remark the secret of continuity is to be revealed. To this I may say that I am glad if everyone finds the above principle so obvious and so in harmony with his own ideas of a line; for I am utterly unable to adduce any proof of its correctness, nor has anyone the power. The assumption of this property of the line is nothing else than an axiom by which we attribute to the line its continuity, by which we find continuity in the line. If space has at all a real existence it is *not* necessary for it to be continuous; many of its properties would remain the same even were it discontinuous. And if we knew for certain that space was discontinuous there would be nothing to prevent us, in case we so desired, from filling up its gaps, in thought, and thus making it continuous; this filling up would consist in a creation of new point-individuals and would have to be effected in accordance with the above principle.

These last few sentences represent an attempt to forestall the probable objection that one cannot simply invent mathematical objects *ex nihilo*. By imagining the completion of a disconnected space, Dedekind was arguing that in fact mathematicians had been doing precisely that for millennia. They had *assumed* that the familiar continuity properties of space were synthetic *a priori* knowledge, as Kant would have said. Dedekind was saying that, on the contrary, nothing whatever was known about physical space *a priori*, and that the space studied in geometry was a creation of the human mind. By arguing that mathematicians had been creating points of a line *ad hoc*, he was justifying a similar creation of irrational numbers.

Having exhibited the geometric principle that had been used in analysis, Dedekind showed in the fourth Section, which is devoted to the creation of irrational numbers, how this same principle could be applied to obtain a concrete representation of the irrational numbers. He defined a *cut* in the rational numbers to be any separation of them into two classes such that every number in the first class is smaller than every number of the second class. He noted that each rational number produces such a cut, in fact two cuts, 'which, however, we shall not look upon as essentially different'. He then showed how to define a cut corresponding to the square root of a non-square positive integer D, by assigning to the second class every positive rational number whose square is larger than D, and all other numbers to the first class. By considering the mapping

$$x \mapsto \frac{x(x^2 + 3D)}{3x^2 + D} = y \tag{1}$$

he showed that the first class had no largest element and the second class no smallest element. (If the square of the positive number x is less than D, then y is also positive and larger than x and has square less than D; if x is positive and its square is larger than D, then y is also positive and smaller than x and has square larger than D.) Dedekind concluded this section by showing that the cuts of the rational numbers form a system with the same order properties as those already listed for the line (assuming the identification of the two cuts produced by a rational number).

The very short fifth Section, treating the continuity of the real numbers, was devoted to proving that cuts made in the system of cuts of the rational numbers did not lead to any new objects, that is, that every cut of the enlarged system was produced by a unique cut in the original system of rational numbers, so that the real numbers thereby created were complete.

The sixth Section (operations with the real numbers) gave an outline of the way in which one could make the cuts of the rational numbers into a number system. Dedekind gave the details only for addition, and left it to the reader to imitate the argument so as to define the other operations of arithmetic, noting that this procedure made it possible at last to prove arithmetically the rules for operating with square roots and other similar rules, which had been taken for granted in the past. He recognized that his definition was somewhat cumbersome and suggested a way of streamlining it by consideration of intervals. By showing that the endpoints of intervals can be simply obtained by cuts, he came very close to formulating explicitly what is now called the *least upper bound* property of the real numbers, which asserts that every set of real numbers that has an upper bound has a smallest upper bound.

In the seventh and last Section, on infinitesimal analysis, Dedekind gave the application to calculus by proving that an increasing bounded function approaches a limiting value. This example and the corresponding proof that the oscillations of a variable approaching a limit must tend to zero were, he thought, sufficient to show how his definition of continuity could be used to derive infinitesimal analysis.

5 RECEPTION OF THE WORK

As is often the case with a difficult problem, once a modicum of clarity has been reached through a rather arduous process, it becomes possible to simplify the entire presentation. Such a simplification, for example, had occurred with the Eudoxan definition of geometric proportion, and a simplification of Dedekind's postulate in the form of the least upper bound axiom was soon introduced into analysis. Dedekind had, in modern terms, provided a model of a structure that has the properties of a number system and at the same time those of a geometric line. That was the kind of object real analysis required. Dedekind's language is revealing; he knew that he was creating new numbers, but he was careful to say that his cuts *corresponded* to the numbers, not that they *were* the numbers:

> Whenever, then, we have to do with a cut [. . .] produced by no rational number, we create a new, an *irrational* number [. . .] which we regard as completely defined by this cut [. . .]; we shall say that the number [. . .] corresponds to this cut, or that it produces this cut. From now on, therefore, to every definite cut there corresponds a definite rational or irrational number [. . .].

One of the burdens of pointing out a common oversight is that people who have been making the oversight fail to see that any difficulty is being overcome. Euclid had shown how to treat proportion geometrically, and Descartes had shown how to represent the product of two lines as a line. Surely combining the work of these two, one would have thought, gave a sufficient basis for treating the real line as a set of numbers. So, at least, believed Lipschitz, a competent mathematician to whom Dedekind sent a copy of his work in 1876. Lipschitz did not see what the fuss was about, and he objected to Dedekind's claims of originality [Scharlau, 1986, 58]:

> I must say that I do not deny the validity of your definition, but that I am nevertheless of the opinion that it differs only in form, not in substance, from what was done by the ancients. I can only say that I consider the definition given by Euclid: *rationem habere inter se magnitudines dicuntur, quae possunt multiplicatae sese mutuo superare* [magnitudes are said to have a ratio if they are capable of exceeding each other when multiplied] and so forth, to be just as satisfactory as your definition. For that reason, I wish you would drop the claim that such propositions as $\sqrt{2}\sqrt{3} = \sqrt{6}$ have never been proved. I think the French readers especially will share my conviction that Euclid's book provided necessary and sufficient grounds for proving these things [. . .]. To quote Jacobi, these questions touch an analyst's heart very deeply, and I only hope you are not angry with me.

Dedekind was not offended, but also not convinced. He replied [Scharlau, 1986, 64–65]:

> I have never imagined that my concept of the irrational numbers has any particular merit; otherwise I should not have kept it to myself for nearly fourteen years. Quite the reverse, I have always been convinced that any well-educated mathematician who seriously set himself the task of developing this subject rigorously would be bound to succeed [...]. Do you really believe that such a proof can be found in any book? I have searched through a large collection of works from many countries on this point, and what does one find? Nothing but the crudest circular reasoning, to the effect that $\sqrt{a}\sqrt{b} = \sqrt{ab}$ because $(\sqrt{a}\sqrt{b})^2 = (\sqrt{a})^2(\sqrt{b})^2 = ab$; not the slightest explanation of how to multiply two irrational numbers. The proposition $(mn)^2 = m^2n^2$, which is proved for rational numbers, is used unthinkingly for irrational numbers. Is it not scandalous that the teaching of mathematics in schools is regarded as a particularly good means to develop the power of reasoning, while no other discipline (for example, grammar) would tolerate such gross offenses against logic for a minute? If one is to proceed scientifically, or cannot do so for lack of time, one should at least honestly tell the pupil to believe a proposition on the word of the teacher, which the students are willing to do anyway. That is better than destroying the pure, noble instinct for correct proofs by giving spurious ones.

Lipschitz was correct in his belief that the French mathematicians would regard Euclid's argument as sufficient. Dedekind, an early enthusiastic proponent of set theory, made the first attempt to derive arithmetic from logic and set theory in his essay *Was sind und was sollen die Zahlen?* (1888) (§47). In this essay Dedekind noted that Jules Tannery had, apparently independently of him, discovered the idea of a cut, but had attributed it to a remark of Joseph Bertrand, to the effect that an irrational number is defined by the set of rational numbers less than and greater than it. To Dedekind, this neglect of detail concealed a fatal flaw, in the form of references to measuring one number by another.

The consensus of mathematical opinion has sided with Dedekind in this debate. It has been considered necessary to lay down very systematically the rules for operating with real numbers, and they are now characterized as an ordered field satisfying the Dedekind postulate (or, equivalently, as a complete Archimedean-ordered field). The powerful influence of these thirty-odd pages is amply attested by a number of facts. The work itself went through five editions in its first half-century. The influence of Dedekind's ideas was so profound that when his collected works were published in 1931, the editors wrote at the end of this essay, 'The argument in this classical work is so well known that we consider it permissible to dispense with commentary'. The method of Dedekind cuts was repeated almost word for word in many standard textbooks of real analysis throughout the 20th century (for example, [Hardy, 1908] and [Rudin, 1953]). To be sure, references to Dedekind cuts have become less common of late; Cantor's approach, which defines a real number as an equivalence class of Cauchy sequences of rational numbers, seems to be more common, since the same argument can be adapted to embed any metric space as a dense subset of a complete space. The late Einar Hille, who worked for a time with Gustav Mittag-Leffler in Stockholm, reported that the latter exploded in exasperation when he told him he had

used Dedekind cuts to define the real numbers [Hille, 1980]. More often than not, authors no longer share the worry that caused Dedekind and Cantor to argue for the possibility of creating new numbers out of the rational numbers. In most of the now-current textbooks one finds the real numbers treated axiomatically, with Dedekind's postulate replaced by the least upper bound axiom. The existence and uniqueness of such an object are generally ignored, as a topic of interest only to philosophers of mathematics, with which the mathematician need not be troubled. However, the recent textbook [Strichartz, 1995] discusses a variety of ways of constructing the real numbers, including Dedekind cuts.

Dedekind's clarity remained a model for other mathematicians to use when creating new objects. One such object, for example, was the set of countable ordinal numbers, that is, ordinal numbers of types I (finite) and II (countably infinite). All attempts to enumerate the second class in an explicit way lead to confusion, and some way of finding a canonical representative of each such ordinals in terms of standard mathematical objects would be highly desirable. The Russian mathematician Nikolai Luzin made many attempts to do so in his unpublished notebooks, and in one note in the archives of the Russian Academy of Sciences he attempted to imitate Dedekind's construction of the real numbers by defining a countable ordinal number to be a well-ordered subset of the rational numbers between 0 and 1 [Cooke, 1993]. Luzin said:

> An irrational number is defined as the symbol for a pair of classes A and B into which the set of rational numbers is decomposed by virtue of some definite rule. Irrational numbers are considered equal if they are symbols for the same pair of classes. After this definition is made, the elementary properties of irrational numbers are established, and finally one can speak of the *totality* of all irrational numbers. Since the work of Russell, a route that is to some extent analogous has become feasible in the theory of *transfinite numbers of second type*. [Then follows the definition of a second-type transfinite ordinal as a well-ordered infinite set of rational numbers between 0 and 1, with order-isomorphic sets identified.] Looking closer, we observe a certain difference between the definition of an irrational number and a transfinite number of second type. Although both definitions have the rational numbers as their point of departure, an irrational number is defined as a pair of classes [...] while a transfinite number of second type is a class whose elements are no longer rational numbers, but *sets* made up of rational numbers.

Thus Luzin used Dedekind's technique as a model for creating transfinite ordinal numbers, and also as the touchstone by which he judged the result and found it wanting. The fact that Luzin thought of it at all is testimony to the extent to which Dedekind's ideas had become a paradigm for creating new numbers.

While Dedekind's ideas have added clarity to mathematical analysis and the philosophy of mathematics, they have been subject to attack from two opposite sides. On the one side, a group of mathematicians with a deep interest in philosophy and logic, the intuitionists, has pointed out that one cannot always decide whether one rational number is larger than another, so that Dedekind's use of ordering to define real numbers does not work in general. For example, it is not known whether the number $(-1)^n$, where n is the quintillionth decimal digit of the positive 15th root of 2, is a positive number; and it is not likely ever

to be known. From the opposite side, there are attacks from practical-minded people of all stripes, who disdain philosophy and deny that the notion of infinite precision implied by a geometric line, which Dedekind was trying to formulate in words, can have any meaning at all. For such people, the 'construction' of the real numbers, whether by Dedekind cuts or otherwise, fits the verdict that Paul Gordan pronounced on one proof of the Hilbert basis theorem: 'This is no longer mathematics, it is theology' [Kowalewski, 1950, 25]. Such a view was forcefully expressed by Norman David Mermin: 'Bridges would not be safer if only people who knew the proper definition of a real number were allowed to design them' [Mermin, 1979].

In the last analysis, the geometric intuition from which the idea of continuity and the idea of an infinitely precise real number arise does not mesh well with the finite, verbal aspect of mathematics that is adapted for logical inference and computation. Yet it is reassuring to know that a precise verbal description of an intuitive geometric idea does exist, to which one can have recourse when intuition becomes doubtful. To have provided that description is the lasting achievement of Dedekind's pamphlet.

BIBLIOGRAPHY

Cantor, G. 1872. 'Ueber die Ausdehnung eines Satzes aus der Theorie der trigonometrischen Reihen', *Mathematische Annalen*, 5, 123–132. [Repr. in *Gesammelte Abhandlungen*, 1932, 92–102.]

Cavaillès, J. 1962. *Philosophie mathématique*, Paris: Hermann. [Contains a French translation of the Cantor–Dedekind correspondence cited below.]

Cooke, R. 1993. 'Arkhiv Luzina', *Istoriko-matematicheskie Issledovaniya*, *24*, 246–255.

Dedekind, R. 1964. *Über die Theorie der ganzen algebraischen Zahlen*, Braunschweig: Vieweg. [Original 1879.]

Ferreirós, J. 1993. 'On the relations between Georg Cantor and Richard Dedekind', *Historia mathematica*, *20*, 343–363.

Ferreirós, J. 1999. *Labyrinth of thought*, Basel: Birkhäuser.

Hardy, G.H. 1908. *A course of pure mathematics*, 1st ed., Cambridge: Cambridge University Press.

Hille, E. 1980. 'In retrospect', *Mathematical intelligencer*, *3*, no. 1, 3–13.

Klein, F. 1926. *Vorlesungen über die Entwicklung der Mathematik im 19. Jahrhundert*, vol. 1, Berlin: Springer-Verlag. [Repr. New York: Chelsea, 1969.]

Kowalewski, G. 1950. *Bestand und Wandel*, München: Oldenbourg.

Mermin, N.D. 1979. Review of *The topological theory of defects*, in *Review of modern physics*, *51*, 591–648.

Noether, E. and Cavaillès, J. (eds.) 1937. *Briefwechsel Cantor–Dedekind*, Paris: Hermann.

Rudin, W. 1953. *The principles of mathematical analysis*, New York: McGraw–Hill.

Scharlau, W. (ed.) 1986. *Rudolf Lipschitz. Briefwechsel mit Cantor, Dedekind, Helmholtz, Kronecker, Weierstrass*, Braunschweig: Vieweg; Wiesbaden: Deutsche Mathematiker-Vereinigung.

Strichartz, R. 1995. *The way of analysis*, Boston: Jones and Bartlett.

CHAPTER 44

JAMES CLERK MAXWELL, *A TREATISE ON ELECTRICITY AND MAGNETISM*, FIRST EDITION (1873)

F. Achard

This comprehensive mathematical treatise on electricity and magnetism led to a slow spread of Maxwell's theory of the electromagnetic field and of his electromagnetic theory of light, first published in 1865. After Hertz's experiments in 1888, it became recognized as the major source of aether and field physics.

First publication. 2 volumes, Oxford: Clarendon Press (hereafter 'CP'), 1873. xxix + 425; xxiii + 444 pages. Print run: 1,500 copies.

2nd edition. 2 volumes (ed. W.D. Niven), CP, 1881. xxxi + 464; xxiii + 456 pages.

3rd edition. 2 volumes (ed. J.J. Thomson), CP, 1891. xxxii + 506; xxiv + 500 pages. [Photorepr. New York: Dover, 1954; New York: Oxford University Press, 1998.]

German translation of the 2nd ed. *Lehrbuch der Electricität und des Magnetismus* (trans. B. Weinstein), 2 vols., Berlin: J. Springer, 1883.

French translation of the 2nd ed. *Traité d'électricité et de magnétisme* (trans. G. Séligmann-Lui, notes by A. Cornu, A. Potier and E. Sarrau), 2 vols., Paris: Gauthier–Villars, 1885–1887. [Photorepr. Sceaux: Jacques Gabay, 1989. Available on the web site of the *Bibliothèque Nationale de France*, as is the *Treatise* (1st ed.).]

Italian translation of the 3rd ed. *Trattato di elettricita e magnetismo* (trans. E. Agazzi), 2 vols., Turin: UTET, 1973.

Russian translation of the 3rd ed. *Traktat ob elektrichestve i magnetizme* (trans. B.M. Mikhailovitch, M.L. Levin), 2 vols., Moscow: Nauka, 1989.

Abridged version

Elementary treatise on electricity (ed. W. Garnett), CP, 1881. 2nd ed. 1888.

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

German translation. *Die Elektrizitat in elementarer Behandlung* (trans. L. Graetz), Braunschweig: Vieweg, 1883.

French translation. *Traité élémentaire sur l'électricité* (trans. G. Richard), Paris: Gauthier–Villars, 1884.

Manuscripts. Parts of the final draft of the *Treatise* (1st ed.), a few folios from earlier drafts and some proofs annotated by Maxwell and P.G. Tait are held in Cambridge University Archives, Cambridge, United Kingdom. The final draft of the *Elementary treatise on electricity* is also there. See also [Maxwell, 1990–2002], vol. 2.

Related articles: Lagrange on mechanics (§16), Fourier (§26), Green (§30), Hamilton (§35), Thomson and Tait (§40), Heaviside (§49), Rayleigh (§45), Kelvin (§58), Hertz (§52), Lorentz (§60), Einstein (§63).

1 EDUCATION AND CAREER

James Clerk Maxwell was born in Edinburgh on 13 June 1831. His father was descended from a long line of Scottish baronets and had been trained as a lawyer. In 1841 he entered the Edinburgh Academy, where he was a contemporary of Peter Guthrie Tait. In 1847 he entered the University of Edinburgh and followed the lectures of the Professor of Natural Philosophy, James David Forbes, and the Professor of Logic, Sir William Hamilton (not to be confused with the inventor of quarternions). In 1850 he left for Cambridge, and in 1854 came second in the Mathematical Tripos (hereafter, 'MT') and first ex-aequo for the Smith's Prize. In addition to his intensive mathematical education, mainly in the hands of the private coach William Hopkins, he probably came under the influence of George Stokes (1819–1903), the Lucasian Professor of Mathematics; and William Whewell, the Master of Trinity College. In 1850 he also made the acquaintance of William Thomson (1824–1907, later Lord Kelvin), the young Professor of Natural Philosophy at Glasgow, who certainly became a model for him.

Maxwell's career as a teacher began at Cambridge in 1855 as Fellow of Trinity College. In the following year he became Professor of Natural Philosophy at Aberdeen. He had to resign in 1860, following changes in the university system, and was an unsuccessful candidate for the Chair of Natural Philosophy at Edinburgh (obtained by Tait). In the same year, however, he was appointed Professor of Natural Philosophy at King's College, London. He left this position in 1865, probably in order to devote himself fully to scientific research. Over the next few years, he spent most of his time at his ancestral home in Scotland, residing in London in the winter months. It was during this period that he wrote his *Treatise on electricity and magnetism*. In 1871 he was appointed to the new Chair of Experimental Physics at Cambridge and the Directorship of the Cavendish Laboratory. He occupied this position until his death on 5 November 1879.

In 25 years of activity, Maxwell published about a hundred scientific papers. While his main claim to fame lies in his work on electromagnetism and the kinetic theory of gases, he was interested in almost all branches of physics, both mathematical and experimental, and especially in mechanics, geometry and optics. In 1857 his essay on the rings of Saturn earned him the Adams Prize at Cambridge, and in 1860 he won the Rumford Medal of the

Royal Society for his work on colour vision. He was made a Fellow of the Royal Society in 1861, and subsequently became a member of several other learned societies. He also published two books aimed at popularizing advanced topics, the *Theory of heat* (1871) and *Matter and motion* (1877), and a very full edition of the unpublished manuscripts of Henry Cavendish on electricity (1879).

2 MATHEMATICAL THEORIES OF ELECTRICITY AND MAGNETISM IN THE FIRST HALF OF THE 19TH CENTURY

One generally thinks of the mathematical treatment of electric phenomena as beginning with two memoirs on electrostatics by Siméon Denis Poisson (1781–1840) in 1812 and 1813. He based his study of conditions of equilibrium for electricity in a conductor on the Newtonian law of force between two (small) electrified spheres established by Charles Coulomb in the 1780s:

$$F \propto \frac{ee'}{r^2}, \tag{1}$$

where F was the force between the spheres, e and e' their respective electric charges, and r the distance between their centres. Using results of P.S. Laplace and J.L. Lagrange on the theory of gravitation, he found a close agreement between his analytic study and experimental observation. Like Coulomb, he assumed that electricity was being composed of two electric fluids. In 1826 and 1827, he also described a mathematical theory of magnetism based on Coulomb's law for magnetic action and the hypothesis that the magnetization of a body results from the separation of two magnetic fluids in the interior of each molecule [Grattan-Guinness, 1990, 496–514, 948–953, 961–965].

Poisson made frequent use of a function V which was later, under the name 'potential', to play an important role in the work of George Green (§30) and C.F. Gauss. The value of V at a point M is determined from the distribution of electricity by the integral formula

$$V = \iiint \frac{\rho}{r}\, dx\, dy\, dz, \tag{2}$$

where ρ is the volume density of the electric charge at a point M' at a distance r from M. The coordinates of the electric force on a unit positive electric were given by

$$E_x = -\frac{dV}{dx}, \qquad E_y = -\frac{dV}{dy}, \qquad E_z = -\frac{dV}{dz}. \tag{3}$$

In the memoir of 1813, he showed that V satisfied the local differential equation, later called 'Poisson's equation', at every point of the space:

$$\frac{d^2V}{dx^2} + \frac{d^2V}{dy^2} + \frac{d^2V}{dz^2} = -4\pi\rho. \tag{4}$$

The phenomenon of the action of an electric current on a magnetized body, discovered by Hans Christian Oersted in 1820, attracted the attention of many experts. In the same

year, André-Marie Ampère (1775–1836) established by experiment the actions of repulsion and attraction between two conducting wires carrying an electric current, and explained all magnetic actions as deriving from this last phenomenon by supposing magnetic bodies to be composed of molecular electric currents. He coined the terms 'electrostatics' and 'electrodynamics' to distinguish the study of forces on bodies carrying electricity at rest and in motion. Ampère's electrodynamic theory, which finally appeared in a memoir of 1826, was based on a new Newtonian formula expressing the interaction between two elements of current:

$$F = ii' \frac{dl\, dl'}{r^2} \left(\sin\alpha \sin\beta \cos\gamma - \frac{1}{2} \cos\alpha \cos\beta \right), \tag{5}$$

where r was the distance between the elements of current, i and i' the intensities of the currents passing through them, dl and dl' are their lengths, and α, β, γ angles expressing their relative orientation. He also established the equality of the magnetic actions exercised by an electric circuit and a magnetic shell occupying a surface bounded by the circuit [Grattan-Guinness, 1990, 917–968].

In 1831 Michael Faraday (1791–1867) established the existence of the phenomenon of electromagnetic induction, but it was not until 1845 that Franz Neumann gave it a mathematical treatment. Appealing to the electrodynamic theory of Ampère and a qualitative law announced by Emil Lenz in 1834, he established, in the case of two closed circuits in relative motion, an expression for the electromotive force of induction as a function of a 'potential' P. This theory established a connection with the electrodynamic theory, since the electrodynamic force on one of the circuits is obtained by differentiating P with respect to the spatial coordinates [Darrigol, 2000, 45–49, 400–401].

In a memoir published in 1846, Wilhelm Weber attempted to unify all of the electric and magnetic phenomena under a Newtonian formula of interaction between two charged particles. To integrate electrodynamic actions and electromagnetic induction, this formula incorporated the first and second derivatives of their distance with respect to time, so as to give:

$$F = \frac{ee'}{r^2} \left[1 - \frac{1}{C^2} \left(\frac{dr}{dt} \right)^2 + \frac{2r}{C^2} \left(\frac{d^2 r}{dt^2} \right) \right], \tag{6}$$

where e and e' were the charges of the two particles, r their distance apart, and C a constant whose importance will appear below. Weber also adopted the hypothesis, proposed by his colleague Gustav Fechner, according to which an electric current is composed of a double flux of positive and negative fluids of equal and opposite rates of flow. Weber was thus able to deduce both Ampère's formula for the interaction between two elements of a circuit and the expression for the electromotive force of induction given by Neumann [Darrigol, 2000, 54–66, 402–405].

3 FARADAY AND THOMSON ON THE NOTION OF FIELD

From 1831 to 1852, Michael Faraday published his 'Experimental researches on electricity and magnetism' in the *Philosophical Transactions of the Royal Society*. These papers

contain not only an impressive series of experimental discoveries, but also a collection of heterodox theoretical concepts on the nature of these phenomena expressed in terms of lines of force and fields.

In 1838, in the 11th series of his 'Experimental researches', Faraday explicitly rejected the idea of an electrostatic action exercised directly at a distance and attempted to prove that electric induction is propagated by contiguous particles of the insulating medium around the bodies (the 'dielectric'). Moreover, the electric charges observed on the surface of the conductors resulted, according to him, not from an accumulation or deficiency of electric fluid, but from the polarization of the dielectric. To substantiate these claims, he showed in particular that the inductive action between the two surfaces of a condenser depends on the nature of the dielectric separating them, a property called its 'specific inductive capacity' [Gooding, 1978].

In 1845, Faraday announced the discovery of an effect of magnetism on polarized light, today called the 'Faraday effect'. A few months later he discovered diamagnetism, and devoted the years that followed to the study of this phenomenon and the development of new theoretical concepts. He disagreed with Weber, for whom diamagnetic bodies possess a polarization opposite to that of paramagnetic bodies. For Faraday, on the other hand, the behaviour of different substances resulted from the *local* tendency of the surrounding space, the 'field', to minimize the perturbation introduced by the bodies and their capacity to conduct the lines of forces more or less well than the surrounding medium. According to case, they move towards the 'strongest' places (where the lines of force were most dense) or the 'weakest' [Gooding, 1981].

While Faraday's experimental discoveries earned him great admiration from his contemporaries, his theoretical ideas were received more with perplexity than enthusiasm. The notions that he developed drew upon visual descriptions of the state of the space surrounding the bodies, which seemed to be purely qualitative and devoid of the precision necessary for a mathematical treatment.

In 1845, the young Thomson published an article in which he showed that Faraday's experiments on the inductive capacity of dielectrics were compatible with the mathematical theory of electrostatics constructed by Poisson. He deduced Faraday's results from the hypothesis of a polarization *at a distance* of the dielectric medium under the influence of the electrified surfaces of the condenser, on the model of reasoning used by Poisson for magnetism. However, in another part of his text, he also asserted that Faraday's physical ideas 'may be made the foundation of a mathematical theory' equivalent to the classical theory. His line of argument appealed to a mathematical 'analogy' between the propagation of heat and an electrostatic system, published in 1842, to envisage a mathematical theory of electrostatics modelled on the theory of heat of Joseph Fourier (§26) [Smith and Wise, 1989, 203–236].

In the years that followed, Thomson developed new concepts and mathematical methods that converged with the theoretical notions of Faraday. First he interpreted the ponderomotive force on an electric or magnetic body as resulting from the tendency of the system to minimize its 'mechanical effect' (or 'potential energy'). Then he showed, using Green's theorem (§30.3), that the potential energy of an electrostatic or magnetic system can be regarded as being distributed over the whole space rather than confined to the surface of the bodies. Finally, he developed a programme for reducing electric and magnetic phenomena

to the mechanical state of the aether in the wave theory of light, especially by means of numerous dynamical analogies and illustrations. In 1856, he even asserted that the Faraday effect proved that magnetism resulted from rotational motions of the aether with axes coinciding with the lines of magnetic force [Smith and Wise, 1989, 237–281, 402–412].

4 MAXWELL AND THE THEORETICAL REFORM OF ELECTROMAGNETISM

Maxwell began his researches on electromagnetism following the completion of his studies at Cambridge in 1854. They were aimed at constructing, at a theoretical level, a unified mathematical theory of electric and magnetic phenomena that would express the methods and ideas of Faraday as an alternative to the theory of Weber. This programme was announced in his first article, 'On Faraday's lines of force', in 1856 [Maxwell *Papers*, vol. 1, 155–229] and continued in two other major texts, 'On physical lines of force' ('PL') in 1861–1862 [*ibidem*, 451–513] and 'A dynamical theory of the electromagnetic field' ('DT') in 1865 [*ibidem*, 526–597]. According to a famous passage in its preface, the *Treatise* (1873) represented the outcome of this programme.

The reference to Faraday in Maxwell's work has often masked the role played there by the texts of Thomson, and above all the search for the continuity with the mathematical theories of Poisson, Ampère and Neumann. Rather than a 'mathematical translation' of Faraday's texts, Maxwell's theoretical programme comprised a reform of those classical mathematical theories within the theoretical framework constructed by Thomson in the course of the previous decade. Maxwell's originality vis à vis Thomson lay in the systematic implementation of this programme, extending it to electrodynamic phenomena and introducing into the mathematical theory notions of Faraday not used by Thomson, notably the duality of quantity and intensity and the electrotonic state.

Maxwell's publications were pervaded by a tension between the problem of treating new analytic expressions as empirically founded and that of associating them with descriptions of the physical process whereby the medium is supposed to propagate electric and magnetic actions. In 1856, he used an analogy between systems of attraction following the inverse-square law and the motion of an incompressible fluid to introduce new mathematical structures, explicitly avoiding the presentation of a 'physical theory' of the phenomena. In 1861, on the other hand, he presented a hypothetico-deductive argument exhibiting a medium composed of molecular vortices as a possible cause of the phenomena. In 1865 he took a middle course in describing a theory based on a set of eight 'general field equations', which were at the same time introduced as 'deduced from experimental facts' and associated with 'dynamical illustrations'.

Moreover, the article of 1856 contained some problems and gaps that later publications attempted to resolve. On the one hand, he gave no account of the connection between the electrostatic and electrodynamic theories since the treatment of the latter was limited to the magnetic effects of closed circuits. On the other hand, the hypothesis of an electromagnetic medium present in a 'so-called vacuum' raised the problem of its co-existence with the aether in the wave theory of light.

The third Part of PL contains a solution of both of these problems. Maxwell first assumed the existence of a new form of electric current consisting of a variation in the electric polarization (or displacement) in a dielectric and deduced a law of magnetic effects

equally applicable to both open and closed circuits. He then gave an argument showing that his electromagnetic medium coincides with the aether in the wave theory of light. He made particular appeal to the close agreement between experimental measurements of the speed of light and a quantity v equal to the ratio $C/\sqrt{2}$, where C is the constant appearing in Weber's formula (6).

In 1865, Maxwell expressed these two solutions in a new form. First, a combination of two of the general field equations implied a new representation of the magnetic effects of closed circuits, according to which the current of conduction in a conducting wire is extended by a current of displacement in the dielectric to form a closed 'total current'. Then he obtained a wave equation from certain field equations from which he again deduced a speed of propagation equal to v. He thus concluded not only that 'light and magnetism are affections of the same substance' but also that 'light is an electromagnetic disturbance propagated through the field according to the electromagnetic laws'. With this last assertion was born the 'electromagnetic theory of light' properly speaking [Siegel, 1991].

5 THE PUBLICATION, FUNCTIONS AND STRUCTURE OF THE *TREATISE*

Around 1860, the rapid development of the telegraphic industry in Britain created an increasing need for knowledge, both theoretical and experimental, of electricity. This is attested by the central role of Thomson in the installation of the submarine telegraphic cable between Britain and the United States, which was completed in 1866 [Smith and Wise, 1989, 649–683]. In 1861, the BAAS created, on Thomson's initiative, a committee charged with defining a standard of electric resistance for use in industry, which Maxwell joined in 1862. From this work he gained the perspective of a precise experimental measurement of v, expressing the ratio of the electromagnetic and electrostatic units of electricity, thereby justifying its conjectured equality with the speed of light [Schaffer, 1992, 1995].

From the middle of the 1860s, several British universities began teaching this new knowledge and scientific practice, creating chairs of experimental physics and associated teaching laboratories [Gooday, 1990]. These innovations also affected the University of Cambridge. In July 1867, it was decided to remodel the MT with a particular view to including the study of electricity, magnetism and heat. Thanks to a donation by the seventh Duke of Devonshire, the Chancellor of the University and a relative of Henry Cavendish, it was decided in 1870 to create a new chair of experimental physics and the now famous Cavendish Laboratory ('CL') [Sviedrys, 1970].

Maxwell played an important part in the reforms at Cambridge. Between 1866 and 1873, he was five times an examiner for the MT and the chief setter of questions on the new subjects. His nomination in March 1871 for the new chair of experimental physics and the directorship of the CL show that he was regarded as the principal British scientific expert on these subjects after Thomson, and also one of the chief architects of the current reforms in Cambridge [Harman, 1995, 33–37].

The publication of the *Treatise on electricity and magnetism* in 1873 was a direct result of these reforms. Maxwell announced his project in 1867, only a few months after the announcement of the reform of the MT, and the book was published in March 1873, just two months after the first session under the new regime. The publication of an advanced

work of reference on the subject was an essential ingredient of the success of the reform at Cambridge [Achard, 1998]. The book was also closely connected with the *Treatise on natural philosophy* ('TNP') by Thomson and Tait (§40). Both books were published by the Clarendon Press ('CP'), publishers to the University of Oxford; the arrangement allowed Thomson and Tait to set electromagnetism aside. Throughout the preparation of his work, Maxwell kept up a correspondence with the professors at Glasgow and Edinburgh [Harman, 1995, 24–33].

This context explains at least two functions of the *Treatise*: to describe the chief instruments and methods of measurement of the phenomena, for the benefit of experimenters and engineers; and to give an account of the sophisticated techniques for the mathematical treatment of electricity and magnetism, mainly for the students of the MT. To these must be added a third function, more familiar to us because of the reference to Faraday in the preface: to promote Maxwell's own theoretical ideas, which were still little known, even in Britain. The contents of his book are summarized in Table 1.

This situation leads us to wonder how Maxwell tried hard to reconcile such different aims in his book. It has recently been shown that students and teachers preparing for the MT could study certain chapters of the treatise without first assimilating the theory of the electromagnetic field. The same was true for engineers and experimenters interested in the techniques of electric and magnetic measurement [Warwick, 2003, 286–317]. As indicated by Maxwell in his preface, the chapters dealing with these various 'numerical' and experimental aspects are placed at the end of each of the four parts of the *Treatise*.

Paradoxically enough, Maxwell's theoretical innovations were mainly concerned with the early chapters, on the 'elementary parts of the theory'. But in conformity with the style of his earlier theoretical papers, he introduced the new ideas progressively, without disrupting the exposition of the main results of the classical mathematical theory.

6 MATHEMATICAL STRUCTURES IN THE *TREATISE*

The title of the preliminary chapter applies only to the first six articles, which are devoted to the dimensional theory of physical quantities in an 'absolute' system of units based on the unity of length, time and mass. The rest of the chapter described ideas and mathematical results regarded by Maxwell as representative of his theory of the field.

In his earlier papers, Maxwell made frequent use of vector functions expressing properties of the electromagnetic field. But he expressed their relations in terms of Cartesian coordinates, showing no predilection for the theory of quaternions as studied and developed by Tait since 1857 (§35.4). In November 1870, during his second spell of work on the book, he declared to Tait his intention to 'leaven [his] book with Hamiltonian ideas'. In the following year, he published an article on the 'mathematical classification of the physical quantities', which contained the essentials of his preliminary chapter.

In the *Treatise*, Maxwell distinguished the 'ideas' favoured by the quaternions from their 'operations and methods' (art. 10). For him, they provided a 'primitive and natural' means of expressing the relations between vectorial entities without recourse to Cartesian axes. Thus he insisted on the distinction between scalars and vectors and reserved a special type of symbol for the latter (the 'German letters'). On the other hand, he totally ignored the affiliation between quaternions and complex numbers.

Table 1. Contents by chapters of Maxwell's book.

The second volume starts with Part III. The titles are those placed at the heads of the chapters; sometimes they differ from those in the table of contents (1st edition, and 2nd edition 1st volume). The numbers given are those of the first articles. In the 2nd edition (1881), discussed in section 11 below, I-1 and I-2 contain some alterations; the chapters I-3, I-4, I-5 and I-9 were entirely rewritten.

Chapter(s)	**Subject or/and 'Title(s)' (first art.)**
Preliminary	'On the measurement of quantities' (1).
I	*Electrostatics*
I-1	'Description of phenomena' (27).
I-2 to I-4	Mathematical theory of electrostatics: 'Elementary mathematical theory of statical electricity' (63), 'Systems of conductors'[1] (84), 'General theorems' (95).
I-5	'Mechanical action between electrified bodies'[2] (103).
I-6 to I-8	Geometrical descriptions of the electrostatic field: 'On points and lines of equilibrium' (112), 'Forms of the equipotential surfaces and lines of induction in simple cases' (117) 'Simple cases of electrification' (124).
I-9 to I-12	Analytical procedures: 'Spherical harmonics' (128), 'Confocal quadric surfaces' (147), 'Theory of electric images and electric inversion' (155), 'Theory of conjugate functions in two dimensions' (182).
I-13	'Electrostatic instruments' (207).
II	*Electrokinematics*
II-1 to II-3	Fundamental phenomena and laws: 'The electric current' (230), 'Conduction and resistance' (241), 'Electromotive force between bodies in contact' (246).
II-4 & II-5	Electrolysis: 'Electrolysis' (255), 'Electrolytic polarization' (264).
II-6 to II-8	Mathematical theory of conduction: 'Linear electric currents' (273), 'Conduction in three dimensions' (285), 'Resistance and conductivity in three dimensions' (297).
II-9 & II-10	'Conduction through heterogeneous media' (310), 'Conduction in dielectrics' (325).
II-11 & II-12	Measurements of electric resistance: 'The measurement of electric resistance' (335), 'On the electric resistance of substances' (359).
III	*Magnetism*
III-1	'Elementary theory of magnetism' (371).
III-2 & III-3	Magnetic notions: 'Magnetic force and magnetic induction' (395), 'Magnetic solenoids and shells' (407).
III-4 to III-6	Magnetic induction: 'Induced magnetization' (424), 'Particular problems in magnetic induction' (431), 'Weber's theory of induced magnetism' (442).
III-7 & III-8	Magnetic observations: 'Magnetic measurements' (449), 'On terrestrial magnetism' (465).

Table 1. (*Continued*)

Chapter(s)	**Subject or/and 'Title(s)' (first art.)**
IV	*Electromagnetism*
IV-1 to IV-4	Electromagnetic phenomena: 'Electromagnetic force' (475), 'Ampère's investigation of the mutual action of electric currents' (502), 'On the induction of electric currents' (528), 'On the induction of a current on itself' (546).
IV-5 to IV-8	Dynamical theory of the electromagnetism: 'On the equations of motion of a connected system' (553), 'Dynamical theory of electromagnetism' (568), 'Theory of electric circuits' (578), 'Exploration of the field by means of the secondary circuit' (585).
IV-9 & IV-11	Fundamental equations: 'General equations of the electromagnetic field' (604), 'Dimensions of electric units' (620), 'On energy and stress in the electromagnetic field' (630).
IV-12 to IV-14	Particular cases: 'Current-sheets' (647), 'Parallel currents' (682), 'Circular currents' (694).
IV-15 to IV-19	Electromagnetic instruments and measurements: 'Electromagnetic instruments' (707), 'Electromagnetic observations' (730), 'Comparison of coils' (752), 'Electromagnetic unit of resistance' (758), 'Comparison of the electrostatic with the electromagnetic units' (768).
IV-20 & IV-21	Electromagnetism and light: 'Electromagnetic theory of light' (781), 'Magnetic action on light' (806).
IV-22 & IV-23	Continental theories of the electromagnetism: 'Ferromagnetism and diamagnetism explained by molecular currents' (832), 'Theories of action at a distance' (846).

[1] 2nd ed.: 'On electrical work and energy in a system of conductors'.

[2] 2nd ed.: 'Mechanical action between two electrical systems'.

Maxwell's usage of quaternions in the text reflects this attitude. Such expressions appear occasionally, usually at the end of an article, to express an important formula initially derived in Cartesian form. More rarely, he indicated the possibility of condensing a long argument by the use of quaternions (for example, in art. 522).

Table 2 shows the operations employed by Maxwell. He never wrote a full quaternion, formed as the sum of a scalar part and a vector part; thus he was already very close to the modern usage in vector analysis, which was introduced later by two readers of the *Treatise*, J.W. Gibbs and Oliver Heaviside ([Crowe, 1967, 127–139]; and compare §35.5 and §49).

Another advantage of the quaternions lay in the fact that they could be used to introduce the operator ∇, whose usefulness in mathematical physics had been demonstrated by Tait. This operator is formally defined as follows (art. 17):

$$\nabla = i\frac{d}{dx} + j\frac{d}{dy} + k\frac{d}{dz}, \tag{7}$$

Table 2. Operations of the quaternion calculus used by Maxwell in the *Treatise*.
α and β are vectors, k is a scalar.

Nature of the operation	**Quaternion notation**	**Vectorial analysis equivalent**
Sum of two vectors.	$\alpha + \beta$	$\alpha + \beta$
Product of a scalar and a vector.	$k\alpha$	$k\alpha$
Scalar part of the product of two vectors.	$S \cdot \alpha\beta$	$-\alpha \cdot \beta$
Vectorial part of the product of two vectors.	$V \cdot \alpha\beta$	$\alpha \times \beta$

Table 3. Operations on scalar and vectorial functions used by Maxwell.
Ψ is a scalar function, σ is a vectorial function.

Name of the operation	**Quaternion notation**	**Vectorial analysis equivalent**
'Slope' of Ψ.	$\nabla\Psi$	$\nabla\Psi$
'Convergence' of σ.	$S \cdot \nabla\sigma$	$-\nabla \cdot \sigma$
'Curl' of σ.	$V \cdot \nabla\sigma$	$\nabla \times \sigma$
'Concentration' of Ψ.	$\nabla^2\Psi$	$-\nabla^2\Psi$

where i, j and k were unit vectors along the three axes of Cartesian coordinates. It can thus be manipulated like a vector. Maxwell went on to list some operations involving ∇ on a scalar or vector function, giving them names that reflect an appropriate geometric property (Table 3).

In what follows, we shall use the notation of modern vector analysis, along with that depicted in Tables 4 and 5 in section 9 below, which summarizes Maxwell's usage in Part IV, ch. 9 of letters to identify the principal equations of the field. The main entities of the field and these equations are listed in Tables 4 and 5, but I shall use them from now on.

Maxwell also made a distinction between two types of 'physical' vector quantities, 'forces' (called 'intensities' in the second edition) and 'flux' (arts. 12–14), constituting a new version of the intensity/quantity duality introduced in 1856. This classification, derived from hydrodynamics, distinguished vector functions respectively expressing a motion of material (the flux through a unit of surface) and the tension that caused it (the gradient of pressure). In electromagnetic theory, it expressed the distinction between, on the one hand, the electric and magnetic forces prevailing at a point of the space and, on the other hand, the electric and magnetic polarizations of the medium or current that they generate at that point. The local numerical relation between a force and the flux it produced depends on the nature of the medium. In the case of an isotropic medium, it is expressed by a relation of vector proportionality (equations (F), (G) and (L)). Finally, Maxwell associated with each type of entity a specific type of mathematical operation. Force was simply integrated along a line to obtain the work effected upon a body. Flux was used in double integrals over a surface to express the quantity that traverses it.

Maxwell went on to describe the properties of line integrals and surface integrals (arts. 16–24). In this last section, he announced two mathematical results that had played a cen-

tral role in the development of his theory since 1856, especially in the expression of integral laws in the form of local equations, namely, Theorems III and IV (arts. 21 and 24), today called 'Ostrogradsky's theorem' and 'Stokes's theorem' respectively. They form a part of a long sequence of results on multiple integrals going back to the beginning of the century [Cross, 1985]. Reverting to the notation used by Tait, Maxwell also expressed these results in terms of quaternions (art. 25).

7 ELECTROSTATICS AND ELECTROKINETICS

The first chapter in the part on electrostatics contained an empirical introduction to the fundamental concepts and laws of the theory. Maxwell introduced the notion of electricity as a measurable physical quantity (arts. 27–34), independent of the hypothesis of electric fluids (arts. 35–37). He stated Coulomb's law in the same way (arts. 38–43) and defined the key entities of electrostatics (arts. 44–50). At the end of the chapter he described the plan of the chapters that follow, along with his 'theory of electric polarization' on the nature of charge and electric current as an alternative to the theory of fluids (arts. 59–62).

In the second chapter the classical electrostatic results were derived from Coulomb's law, introducing occasionally some of the ideas of field theory. Thus Maxwell defined the notion of electric displacement as 'the quantity of electricity which is forced in the direction of [the electromotive force] across a unit of area', and wrote down equation (F) (art. 68). He also gave Poisson's equation in generalized form:

$$\frac{d}{dx}\cdot K\frac{dV}{dx}+\frac{d}{dy}\cdot K\frac{dV}{dy}+\frac{d}{dz}\cdot K\frac{dV}{dz}+4\pi\rho=0, \tag{8}$$

incorporating the coefficient K of capacity of the medium derived from experiments of Faraday (art. 83). This equation is equivalent to equation (J) if one assumes equations (F) and (3).

The fourth chapter was chiefly devoted to two theorems, attributed to Green and Thomson respectively, which Maxwell interpreted physically in accordance with the field theory. He also stated them in the generalized form of equation (8). In the fifth chapter, he set out to *explain* the interactions between electrified fields by a distribution of stress from the surrounding medium rather than by a direct action at a distance and to emphase his agreement with the writings of Faraday (arts. 105–109). Finally, he admitted to not having taken the next step in the search for a physical explanation of the phenomena: to account for this distribution of stress in the medium in terms of 'mechanical considerations' (arts. 110–111).

The 'theory of electric polarization' described at the end of the first and fifth chapters (arts. 60–62, 111) asserted, in conformity with ideas put forward by Faraday, that electric charge manifests the discontinuity between the polarized state of a dielectric and the non-polarized state of a conductor. It also stated that 'the motions of electricity are like those of an incompressible fluid'. This means that electric current always forms closed curves in accordance with the equation

$$\nabla\cdot\mathbf{C}=0 \tag{9}$$

(which follows from equation (E)). Moreover, Maxwell emphasized that the 'total' electric current is composed of the 'ordinary' current of conduction, which predominates in conductors, and that the variation of electric displacement, which predominates in dielectrics (whence equation (H)). Finally, he stated that the passage of electricity through a medium generated a state of constraint that loosens and reforms at a frequency more or less rapid depending on the nature of the medium. It is this discontinuity of states of constraint at the surface separating two media that is manifest in the electric charge [Buchwald, 1985, 20–40].

The second Part, on 'electrokinematics', was essentially an exposition of the classical mathematical theory of electric current. Maxwell treated electric current as a phenomenon that is empirically observed and quantitatively measured by a galvanometer (art. 240), without any consideration of its physical nature, and mentions Ohm's law (which corresponds locally to equation (G)) and the Joule effect (arts. 241–242). In Chapter 10, the expression (H) of the total current in a dielectric, as consisting of a current of conduction and a variation of displacement, was used to give an account of the phenomenon of electric absorption (arts. 328–334).

8 MAGNETISM AND ELECTROMAGNETISM

The third Part was taken up mainly with a description of the classical theory of magnetism. The early chapters nevertheless contained an introduction to certain ideas and results belonging to field theory, chiefly in the appearance of 'magnetic induction' as a complement of the classical notion of 'magnetic force'. Following Thomson, Maxwell first defined magnetic induction as the magnetic force exercised on a unit magnetic pole lying in an infinitesimal circular cavity. In this case, magnetic induction was expressed in terms of magnetic force by equation (D) (art. 399). He thus deduced, in particular, that the integral of the surface of magnetic induction over a closed surface is always equal to zero (art. 402). Thus

$$\nabla \cdot \mathbf{B} = 0. \tag{10}$$

He also defined a vector **U**, called the 'potential vector' of magnetic induction, such that, for any surface (S) bounded by a closed curve (C):

$$\iint_{(S)} \mathbf{B} \cdot \mathbf{dS} = \oint_{(C)} \mathbf{U} \cdot \mathbf{dl}. \tag{11}$$

He then used Stokes's theorem to deduce the local equation (A) (art. 405). In the fourth chapter, Maxwell suggested that, according to 'Faraday's method', magnetic induction represents the polarization of a medium under the action of a magnetic force and is expressed, in an isotropic medium, by equation (L) (art. 428).

In the early chapters of the fourth part Maxwell tackled in turn the magnetic effects of electric currents and the phenomenon of induction. But he specifically emphasized the contrast between the second chapter, on the theory of Ampère, and the first and third chapters, which were dominated by 'Faraday's method', that is, the field theory (arts. 493, 502,

528–529). While Ampère's theory proceeded from an initial decomposition of the system, the action on a circuit being calculated as the sum of the actions of each element of current, Faraday's method began with the system taken as a whole, the electromagnetic laws being expressed in terms of properties of the global field. This contrast between mathematical methods reflected that between the physical hypotheses.

In the first chapter, the equivalence of the magnetic effects of an electric circuit and a magnetic sheet enabled Maxwell to express the action on an electric circuit placed in an arbitrary magnetic field (arts. 489–492). At the end of the chapter, he stated a law bearing on the distribution of the magnetic force as a function of that of the electric currents: 'the line-integral of the magnetic force' round a closed curve (C) is equal to $4\pi i$, where i denoted the electric current that flows through an arbitrary surface (S) bounded by (C) (arts. 498–499). That is,

$$\oint_{(C)} \mathbf{H} \cdot \mathbf{dl} = 4\pi i. \tag{12}$$

This law was expressed locally by equation (E). Maxwell stated also that the electric current considered here is made up of both a variation of electric displacement and a current of conduction, so that **C** satisfies equation (H) already used in the second part.

In the third chapter, Maxwell appealed to the work of Faraday and to a series of experiments devised by the Italian physicist Riccardo Felici to state the law of electromagnetic induction: 'the total electromotive force acting around a circuit at any instant is measured by the rate of decrease of the number of lines of magnetic force which pass through it' (arts. 536–541). This last notion was also called 'the magnetic induction through the circuit'. This yields the analytic form of the law of induction,

$$e = -\frac{d}{dt}\left(\iint_{(S)} \mathbf{B} \cdot \mathbf{dS}\right), \tag{13}$$

where (S) was a surface bounded by a closed circuit (C). Although Maxwell attributed the discovery of this law of induction to Faraday, he mentions briefly its 'convergence' with the mathematical theories of induction developed by Neumann, then by Helmholtz, Thomson and Weber (arts. 542–545).

9 THE DYNAMICAL THEORY OF ELECTROKINETIC PHENOMENA, AND THE GENERAL EQUATIONS OF THE ELECTROMAGNETIC FIELD

The fourth chapter of Part IV contains a consideration of the similarities and differences between the phenomenon of self-induction of an electric current and the inertia (or 'momentum') of a fluid in motion in a tube. Maxwell concluded by proposing to deduce the principal structure of the theory of electricity from a dynamical hypothesis that stated that the phenomena were produced by a connected system in motion lying both in the surrounding space and in the conducting bodies (art. 552). According to him, the dynamical theory of Lagrange (§16) made it possible to avoid any more detailed hypothesis on the nature of the motions of this system.

The fifth chapter described the fundamental relations of the Lagrangian theory adapted to the needs of dynamical arguments by Hamilton, then by Thomson and Tait (§40). In the sixth chapter, the state of a system of electric circuits was expressed in terms of variables of the following two types (arts. 568–570). Firstly, the ordinary 'mechanical' variables (x_i) described the form and relative position of the circuits. Secondly, the 'electric' variables (y_i) expressed the position of the electricity in motion in the circuits, and their derivatives $(\dot{y}_i)$ with respect to time gave the intensities of the electric current. The kinetic energy of the system is then the sum of three quadratic functions (art. 571): the ordinary kinetic energy T_m, which depends only on the motions of the circuits,

$$T_m = \frac{1}{2}\sum_i A_i \dot{x}_i^2 + \sum_{i<j} B_{ij}\dot{x}_i\dot{x}_j; \tag{14}$$

the electrokinetic energy T_e, which depends only on the electric currents in the circuits,

$$T_e = \frac{1}{2}\sum_i L_i \dot{y}_i^2 + \sum_{i<j} M_{ij}\dot{y}_i\dot{y}_j; \tag{15}$$

and a third term T_{me} which depended upon the products of one mechanical variable and one electric variable,

$$T_{me} = \sum_{i,j} C_{ij}\dot{x}_i\dot{y}_j. \tag{16}$$

Maxwell also showed that the coefficients depended only on the mechanical variables (art. 572). Finally, a series of experiments allowed that the term T_{me} be regarded as negligible (arts. 574–577).

In subsequent chapters, Maxwell considered only phenomena depending on the electrokinetic energy T_e. He also defined the concept of the *electrokinetic momentum p_i associated with each circuit* (A_i) by setting

$$p_i = \frac{dT_e}{d\dot{y}_i} = L_i\dot{y}_i + \sum_{k\neq i} M_{ki}\dot{y}_k. \tag{17}$$

Using formulae from the fifth chapter, he then expressed the external forces applied to the system by differentiating the electrokinetic energy (arts. 579–580) and applied these results in the case of a system compound of two circuits (arts. 581–584). By differentiating with respect to y_i, Maxwell obtained an expression for the electromotive force Y_i' applied to a circuit (A_i) and not compensated by the resistance of the circuit (art. 579). That is,

$$Y_i' = \frac{d}{dt}\left(\frac{dT_e}{d\dot{y}_i}\right) - \frac{dT_e}{dy_i} = \frac{dp_i}{dt} \tag{18}$$

with

$$Y_i' = E_i - R_i\dot{y}_i, \tag{19}$$

where E_i is the external electromotive force and R_i the resistance of (A_i). According to Maxwell, the internal electromotive force e_i of induction corresponds to the opposite of this expression. Thus, in the case of two circuits, the electromotive force of induction in the circuit (A_2) is given by

$$e_2 = -\frac{dp_2}{dt} = -\frac{d}{dt}(M\dot{y}_1 + L_2\dot{y}_2). \tag{20}$$

Likewise, the external mechanical forces applied to the circuits were obtained by differentiating T_e with respect to the mechanical variables x_i (art. 580):

$$X'_j = \frac{d}{dt}\frac{dT_e}{d\dot{x}_j} - \frac{dT_e}{dx_j} = -\frac{dT_e}{dx_j}. \tag{21}$$

The electrodynamic force X_j on a circuit appeared as the opposite of this expression, and in the case of two rigid circuits he obtained the formula

$$X_j = \dot{y}_1\dot{y}_2\frac{dM_{12}}{dx_j}. \tag{22}$$

In the eighth chapter, Maxwell used the electrokinetic moment p of a circuit (C) to define two vector functions **B** and **U**, *the electrokinetic momentum at a point*, by the equations

$$p = \oint_{(C)} \mathbf{U}\cdot\mathbf{dl} = \iint_{(S)} \mathbf{B}\cdot\mathbf{dS}. \tag{23}$$

He immediately identified these new vectorial entities with the notions of magnetic induction and vector-potential respectively, which were introduced in Part III and were related by the local equation (A) (art. 592). Maxwell finally derived local expressions for the electromotive force of induction and the mechanical force as functions of **U** and **B**, and thus obtained equation (B) for the electromotive force in a conductor moving with a speed **v** and equation (C) for the mechanical force on a conductor traversed by a current of density **C** (arts. 595–603).

The ninth chapter covered the 'general equations of the electromagnetic field', already introduced in earlier parts of the treatise. Without any recourse to the earlier dynamical reasoning, Maxwell then stated equations from (D) to (L) in turn (arts. 605–614). Finally, he obtained an expression for the vector-potential as a function of the distribution of the electric currents according to Ampère's theory of magnetism and under the following condition on **U**, today called a 'gauge condition' (arts. 615–617):

$$\nabla\cdot\mathbf{U} = 0. \tag{24}$$

The principal entities and field equations collected at the end of the chapter are shown in Tables 4 and 5.

Chapter 11 contained expressions for the three types of energy: electrostatic, magnetic and electrokinetic (arts. 630–636). Appealing to the theory of Ampère, Maxwell proposed

Table 4. The principal quantities of the electromagnetic field (art. 618). Vectors are denoted by bold letters, instead of the German letters used by Maxwell.

Vectors	
Electromagnetic momentum at a point (or potential-vector).	**U** (F, G, H)
Magnetic induction.	**B** (a, b, c)
Total electric current.	**C**(u, v, w)
Electric displacement.	**D** (f, g, h)
Electromotive force.	**E** (P, Q, R)
Mechanical force.	**F** (X, Y, Z)
Velocity of a point.	**v** $(\dot{x}, \dot{y}, \dot{z})$
Magnetic force.	**H** (α, β, γ)
Intensity of magnetization.	**I** (A, B, C)
Current of conduction.	**R** (p, q, q)
Scalars	
Electric potential.	Ψ
Magnetic potential.	Ω
Electric density.	e
Density of magnetic 'matter'.	m
Physical properties of the medium (isotropic media)	
Conductivity for electric currents.	C
Dielectric inductive capacity.	K
Magnetic inductive capacity.	μ

to regard the energy of the field as divided over the whole space into two fundamental forms (arts. 637–638): electrostatic (potential) energy,

$$W = \frac{1}{2} \iiint (\mathbf{D} \cdot \mathbf{E})\, d\tau; \tag{25}$$

and electrokinetic (kinetic) energy,

$$T = \frac{1}{8\pi} \iiint (\mathbf{B} \cdot \mathbf{H})\, d\tau. \tag{26}$$

The chapter ended with a consideration of the possibility of explaining magnetic action by a state of stress in the surrounding medium, in line with the corresponding study of electrostatic action in part I, chapter 5 (arts. 639–646).

10 THE ELECTROMAGNETIC THEORY OF LIGHT

As in 1865, the main argument in favour of the electromagnetic theory of light lay in deriving from the general field equations (A), (B), (E), (F), (G), (H) and (L) an expression which reduced, in the case of a dielectric medium, to the following wave equation for the vector potential **U** (arts. 783–784):

Table 5. The general equations of the electromagnetic field (art. 619).

Name of the equation (when given)	**Expression in vectorial analysis**	
Equation of magnetic induction.	$\mathbf{B} = \nabla \times \mathbf{U}$	(A)
Equation of electromotive force.	$\mathbf{E} = \mathbf{v} \times \mathbf{B} - \dot{\mathbf{U}} - \nabla\Psi$	(B)
Equation of mechanical force.	$\mathbf{F} = \mathbf{C} \times \mathbf{B} - e\nabla\Psi - m\nabla\Omega$	(C)[1]
Equation of magnetization.	$\mathbf{B} = \mathbf{H} + 4\pi\mathbf{I}$	(D)
Equation of electric currents.	$4\pi\mathbf{C} = \nabla \times \mathbf{H}$	(E)
Equation of electric displacement.	$\mathbf{D} = \frac{1}{4\pi} K\mathbf{E}$	(F)
Equation of the current of conduction.	$\mathbf{R} = C\mathbf{E}$	(G)
Equation of the total current.	$\mathbf{C} = \mathbf{R} + \dot{\mathbf{D}}$	(H)
	$e = \nabla \cdot \mathbf{D}$	(J)[2]
Equation of induced magnetization.	$\mathbf{B} = \mu\mathbf{H}$	(L)
	$m = -\nabla \cdot \mathbf{I}$	
	$\mathbf{H} = -\nabla\Omega$	

[1] In the third edition, this equation was rewritten

$$\mathbf{F} = \mathbf{C} \times \mathbf{B} + e\mathbf{E} - m\nabla\Omega$$

following a correction proposed by G.F. FitzGerald in 1883.

[2] This expression is consistent with the Cartesian expression given in art. 612 and with similar ones used elsewhere in the *Treatise* (e.g. art. 82); but it is the opposite of the quaternion expression given in art. 619. The consistent quaternion expression is $e = -S \cdot \nabla\mathbf{D}$.

$$K\mu \frac{d^2\mathbf{U}}{dt^2} - \nabla^2\mathbf{U} = \mathbf{0}. \tag{27}$$

Maxwell thereby deduced the existence of 'electromagnetic disturbances' whose speed of propagation was given by

$$V = \frac{1}{\sqrt{K\mu}}. \tag{28}$$

In the case of air, he showed that this number coincided with v, 'the number of electrostatics units of electricity in one electromagnetic unit'.

To establish that 'light is an electromagnetic disturbance', Maxwell compared the various experimental measurements of v, given in the preceding chapter, with those of the speed V_L of light (Table 6). He concluded that his theory, which implied that these two quantities were equal, 'is certainly not contradicted by the comparison of these results such as they are' (arts. 786–787).

As in 1865, Maxwell claimed that light is a special kind of electromagnetic perturbation, whose law of propagation is expressed by (27). This reasoning allowed him to justify a strong thesis with minimal hypotheses, since it rests solely on the equality of v and V_L, along with an acceptance of the general equations of the electromagnetic field. Although,

Table 6. Comparison of the ratio of electric units with the velocity of light (art. 787).

Velocity of light (metres/second)		**Ratio of electric units (metres/second)**	
Fizeau	314 000 000	Weber	310 740 000
Aberration, &c	308 000 000	Maxwell	288 000 000
Foucault	298 360 000	Thomson	282 000 000

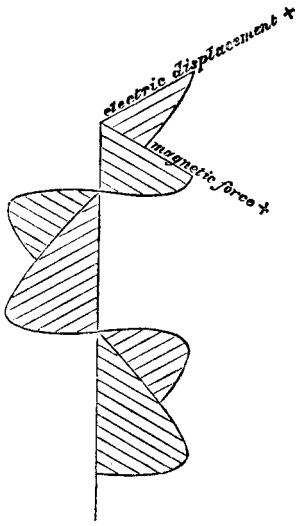


Figure 1. Image of an electromagnetic wave (art. 791). It represents 'the values of the magnetic force and of the electromotive force at a given instant in different points of the ray [...] for the case of a simple harmonic disturbance in one plane'. The 'magnetic and electric disturbances' are transverse to the direction of propagation and perpendicular to each other. 'This corresponds to a ray of plane-polarized light'.

for Maxwell as for his contemporaries, the idea of propagation of 'electromagnetic disturbances' required the acceptance of the hypothesis of a material support, he made no further assumptions about its configuration.

The chapter ended by suggesting various paths that experimental research might take in order to confirm the electromagnetic nature of light by studying the correlation between the optical and electric properties of various substances, such as relationships between the refractive index and inductive capacity of transparent dielectric media (arts. 788–789), the conductivity and opacity of media (arts. 798–805), and also the optical and electric properties of crystalline media (arts. 794–797). Maxwell also presented a study of plane waves and polarized light (arts. 790–791). He showed that, in the case of a plane, the electric and magnetic forces were perpendicular to the direction of propagation of the wave, agreeing with the hypothesis that light waves vibrate transversely (see Figure 1).

Maxwell did not discuss how electromagnetic perturbations might be produced by an electric device, suggesting that he saw no easy way of doing this. Moreover, by choosing in his *Treatise* to derive a wave equation involving the 'electrokinetic moment' **U** rather than the magnetic force **H** as in DT, he seems to be exercising a preference for this entity. This choice introduced a weakness into his reasoning, since to obtain equation (27) he had to assume that Poisson's equation can also be invoked in the non-stationary case of a phenomenon of propagation (art. 783). As remarked by G.F. FitzGerald (1851–1901) in 1890, this involved the assumption that electrostatic potential propagated itself instan-

taneously. It then turned out that this choice was closely connected with the choice of a 'gauge condition' on the vector-potential [Hunt, 1991, 117–118].

In the next chapter, Maxwell treated the phenomenon of magnetic action on light. Inspired by PL, he tried to account for the Faraday effect by supposing that a medium subjected to a magnetic influence was the seat of molecular vortices whose axes of rotation coincide with the lines of magnetic force. By adding a term to the expression for the kinetic energy of the system and using an equation of Lagrange given in the fifth chapter, he arrived at an expression for the angle of rotation of the plane of polarization of the light as a function of various parameters, including the intensity of the magnetism and the length of the ray of light passing through the medium (arts. 822–831). As Maxwell was to remark in 1879, this line of reasoning constituted a 'hybrid' theory, since it explains the magneto-optical phenomenon by supposing that light and magnetism comprise motions of the aether that interact according to mechanical laws rather than by directly invoking the electromagnetic theory of light.

11 THE MAXWELLIANS

Maxwell published no theoretical articles on electromagnetism after 1873. But in 1872 the CP suggested that he write a less mathematical work on electricity and magnetism based on his great treatise, in the style of Thomson and Tait in their *Elements of natural philosophy*. This task, begun around 1874, remained unfinished at the time of his death in November 1879. As well as this, the CP advised him in 1877 that, thanks to the popularity of the *Treatise*, it was already time to start work on a second edition. Table 1 above indicates the changes that he envisaged at that time. Both works were finally edited in 1881 by two of Maxwell's collaborators at Cambridge, W.D. Niven and William Garnett. The unfinished manuscript of the *Elementary treatise on electricity* was completed using sections from the great *Treatise* to cover the material in the first volume (see the publication history at the head of this article).

Maxwell's presence at Cambridge did not, however, immediately result in the formation of a school of research in electromagnetism based on his ideas. Neither his lectures on electricity and magnetism, aimed at students of the MT and the Natural Science Tripos, nor his experimental research conducted at the CL, continuing the metrological work begun by the committee of the BAAS, were chiefly concerned with his field theory. Some graduates, especially those motivated by theoretical research, progressively assimilated the subtleties of the field by following intercollegiate courses delivered by W.D. Niven [Warwick, 2003, 286–356].

Speaking generally, the *Treatise* probably owed its commercial success in Britain to its value in mathematical and experimental education rather than to a rapid acceptance of its theoretical innovations. As early as 1865, William Thomson had reservations about Maxwell's theory, and publicly opposed it in his Baltimore lectures in 1884 (§58). In his review of the *Treatise*, Tait [1873] remarked on the opposition to the idea of action at a distance and the thesis of a 'connection between radiation and electric phenomena', but he ignored essential features such as the physical theories of electric charge and of the displacement current. This was true also of George Chrystal's review of the second edi-

tion [Chrystal, 1882]. Nevertheless, the success of the work certainly contributed to the dissemination of the field theory.

From the end of the 1870s, several dozen 'Maxwellians' began to publish their research on electromagnetism. Many of these came from MT and the CL, notably Joseph John Thomson (1856–1940) and John Henry Poynting (1852–1914), who were both students of Niven. But some of them were related to other British (and American) universities: notably FitzGerald, former student and professor at Trinity College, Dublin; and Oliver Lodge (1851–1940), former student of University College London and professor of physics at Liverpool. The most remarkable case is certainly that of Oliver Heaviside (1850–1925), self-taught by reading works such as the TNP and, above all, Maxwell's *Treatise* (§49).

One of the first extensions of Maxwell's work lay in explaining electric and magnetic phenomenon by a mechanical state of the aether. Some authors, such as Lodge, presented mechanical models to illustrate the production of certain phenomena by the aether [Hunt, 1991, 73–104]. Another direction, particularly favoured by the physicists at Cambridge, consisted of using the Lagrangian theory and the principle of least action to establish a mechanical basis for the phenomena without stating precise hypotheses on the arrangement of the mechanism supposed to produce them [Buchwald, 1985, 54–64].

Although FitzGerald worked on the mechanical models and discussed their merits with Lodge, he also used Lagrangian methods, notably in an influential article published in 1879. A correspondence between Maxwell's theory and the optical theory of MacCullagh led to the incorporation in the electromagnetic theory of light, not only of the phenomena of reflection and refraction of light, but also of the magneto-optical phenomena of Faraday, and of John Kerr discovered in 1876 [Hunt, 1991, 15–23; Buchwald, 1985, 73–129].

From 1879, Lodge sought to confirm experimentally the electromagnetic theory of light by producing light waves by purely electric means. After first asserting the fruitlessness of such an exercise, FitzGerald retracted in 1882 and thought about conditions necessary to produce observable electromagnetic waves. Lodge announced the production of electromagnetic waves in electric wires in 1888, but his triumph was eclipsed by the experiments of Hertz [Hunt, 1991, 24–48, 146–151].

Finally, in 1884, Poynting presented his famous 'theorem'. He deduced from the fundamental equations of field theory a mathematical equality interpreted as expressing that the variation of the energy contained in a given volume per unit time is due to the flux of energy across the surface that bounds it. According to this interpretation, during the passage of current in an electric wire, energy moves, not along the wire, perpendicular to it from the dielectric to the conductor, where it is converted into heat. This interpretation favours Maxwell's ideas on electric current rather than the traditional image of an imponderable fluid in motion [Buchwald, 1985, 41–54; Hunt, 1991, 109–114].

12 THE EXPERIMENTS OF HERTZ AND THEIR IMPACT

Maxwell's treatise was also read in continental Europe. Towards the end of the 1870s, it aroused sufficient interest to warrant translations into German and French, which appeared in 1883 and 1885 respectively. While the treatise disseminated the existence of Maxwell's theoretical ideas, it seems that they did not exercise a strong influence on the majority of

continental research into electromagnetism up to the middle of the 1880s. A theoretical article of Helmholtz published in 1870, even before the appearance of the *Treatise*, constitutes the main and most remarkable exception.

Helmholtz presented a theory of electromagnetic action covering the case of open circuits and taking account of the absence of experimental data on this situation. An expression for the vector potential involved a parameter k whose value expressed the various possible options, three different values corresponding to the theories of Neumann, Maxwell and Weber respectively. All the mathematical laws in Maxwell's theory can be recovered by supposing that space is infinitely polarizable [Buchwald, 1985, 177–186].

In July 1879, Helmholtz worked on a problem for a Prize of the Berlin Academy that was explicitly aimed at testing the validity of Maxwell's prediction about open circuits [Darrigol, 2000, 233–234]. From that date, he encouraged a brilliant student, Heinrich Hertz (1857–1894), to seek a solution of these problems. Hertz provisionally abandoned this project, but in 1884 he published a theoretical account affirming the superiority of Maxwell's theory.

In 1886, Hertz began some experimental research that led to an impressive series of articles published between 1887 and 1889. First he described a device capable of producing extremely rapid electric oscillations and detecting the electromotive forces that they generate. He then emphasized the electrodynamic effect generated by the variable polarization of a dielectric. Then he showed that the propagation of this effect in an electric wire or in air generated progressive or stationary waves according to the experimental configuration. In the latter case, the disposition of the sinks and nodes enabled him to show that the speed of propagation in air is close to that of light. He concluded this study by describing the spatial distribution of waves, with emphasis on their reflection and refraction [Buchwald, 1994].

These experiments, soon repeated by many physicists, assured the triumph of Maxwell's theory and its adherents in Britain. They also had the effect of drawing attention to the work of Heaviside [Hunt, 1991, 158–168]. Inspired in part by this work, Hertz proposed a new formulation of Maxwell's theory in 1890. He also studied the implications of field theory in the rest of physics by developing his ideas on the foundations of mechanics (§52).

In Germany, there was a spectacular surge of interest in Maxwell's theory. In the course of the following decade, several German authors put forward accounts of electromagnetic field theory which abandoned the traditional representations of charge and electric current generally in favour of an agnostic attitude [Darrigol, 2000, 253–262].

13 LARMOR AND THE NOTION OF ELECTRON

In 1894, John Joseph Larmor (1857–1942) published a theoretical article on magnetism that formed one of the sources heralding the advent of the modern notion of electron (see §60 on Lorentz). Although originally conceived as an extension of Maxwell's theory, it diverged in certain central aspects from the Maxwellian approach.

Larmor's theory consisted of supposing that matter is made up of 'isolated singularities of aether', carrying an elementary positive or negative charge, which he called 'electrons'. He conceived of electric current as being formed by the convection of these electrons, and

of properties of matter as resulting from their disposition. Thus, not only did Larmor abandon the Maxwellian representations of electric charge and current, but he also renounced at the same time the macroscopic approach characteristic of Maxwell's work. However, Larmor's theory did preserve the idea that electric and magnetic actions were propagated via the aether [Buchwald, 1985, 131–173; Hunt, 1991, 209–239; Darrigol, 2000, 332–343].

Larmor published the second and third parts of his article in 1895 and 1897. Probably influenced by the work of Lorentz, which he had discovered in the meantime, he developed his theory of electrons to explain the principal magneto-optical phenomena (the Faraday, Kerr and Zeeman effects) and the optical paradoxes of bodies of motion. Finally, he gave a synthetic account of his theory in *Aether and Matter* (1900), which exercised in the next decade an important influence on electromagnetic research in Britain, especially in Cambridge [Warwick, 2003, 357–398].

BIBLIOGRAPHY

Achard, F. 1998. 'La publication du *Treatise on Electricity and Magnetism* de James Clerk Maxwell', *La revue de synthèse*, *119*, 511–544.

Buchwald, J.Z. 1985. *From Maxwell to microphysics*: *aspects of electromagnetic theory in the last quarter of the nineteenth century*, Chicago: University of Chicago Press.

Buchwald, J.Z. 1994. *The creation of scientific effects*: *Heinrich Hertz and electric waves*, Chicago: University of Chicago Press.

Campbell, L. and Garnett, W. 1882. *The life of James Clerk Maxwell*, London: Macmillan. [Repr. New York: Johnson Reprint, 1969. 2nd ed. 1884.]

Chrystal, G. 1882. 'Clerk Maxwell's "Electricity and Magnetism" ', *Nature*, *25*, 237–240. [Review of the second edition and of the *Elementary treatise on electricity*.]

Cross, J.J. 1985. 'Integral theorems in Cambridge mathematical physics, 1830-55', in P.M. Harman (ed.), *Wranglers and physicists*, Manchester: Manchester University Press, 112–148.

Crowe, M.J. 1967. *A history of vector analysis*: *the evolution of the idea of a vectorial system*, Notre Dame: University of Notre Dame Press.

Darrigol, O. 2000. *Electrodynamics from Ampère to Einstein*, Oxford: Oxford University Press.

Gooday, G. 1990. 'Precision measurement and the genesis of physics teaching laboratories in Victorian Britain', *British journal for the history of science*, *23*, 25–51.

Gooding, D. 1978. 'Conceptual and experimental bases of Faraday's denial of electrostatic action at a distance', *Studies in the history and philosophy of science*, *9*, 117–149.

Gooding, D. 1981. 'Final steps to the field theory: Faraday's study of magnetic phenomena, 1845–1850', *Historical studies in the physical sciences*, *11*, 232–275.

Grattan-Guinness, I. 1990. *Convolutions in French mathematics*, *1800–1840*, 3 vols., Basel: Birkhäuser; Berlin: Deutscher Verlag der Wissenschaften.

Harman, P.M. 1995. 'Introduction' to [Maxwell, 1990–2002], vol. 2, 1–37.

Hunt, B.J. 1991. *The Maxwellians*, Ithaca: Cornell University Press.

Maxwell, J.C. *Papers*. *Scientific papers* (ed. W.D. Niven), 2 vols., Cambridge: Cambridge University Press. [Repr. Paris: Hermann, 1927; New York: Dover, 1952.]

Maxwell, J.C. 1990–2002. *Scientific letters and papers* (ed. P.M. Harman), 3 vols., Cambridge: Cambridge University Press. [Vol. 1 1846–1862, vol. 2 1862–1873, vol. 3 1874–1879.]

Schaffer, S. 1992. 'Late Victorian metrology and its instrumentation: a manufactory of Ohms', in R. Bud and S.E. Cozzens (eds.), *Invisible connections*: *Instruments, institutions and science*, Bellingham: SPIE Optical Engineering Press, 23–56.

Schaffer, S. 1995. 'Accurate measurement is an English science', in N.M. Wise (ed.), *The values of precision*, Princeton: Princeton University Press, 135–172.

Siegel, D.M. 1991. *Innovation in Maxwell's electromagnetic theory*: *Molecular vortices, displacement current, and light*, Cambridge: Cambridge University Press.

Smith, C. and Wise, M.N. 1989. *Energy and Empire*: *A biographical study of Lord Kelvin*, Cambridge: Cambridge University Press.

Sviedrys, R. 1970. 'The rise of physical science at Victorian Cambridge', *Historical studies in the physical sciences, 2*, 127–151.

Tait, P.G. 1873. 'Clerk-Maxwell's Electricity and Magnetism', *Nature, 7*, 478–480. [Review of the first edition.]

Warwick, A. 2003. *Masters of theory*: *Cambridge and the rise of mathematical physics*, Chicago: University of Chicago Press.

CHAPTER 45

J.W. STRUTT, THIRD BARON RAYLEIGH, *THE THEORY OF SOUND*, FIRST EDITION (1877–1878)

Ja Hyon Ku

The first comprehensive and systematic mathematical treatise on sound, this book opened the era of modern acoustics. New methods and notions introduced in it are useful today in physics and engineering as well as in acoustics.

First publication. 2 volumes, London: Macmillan, 1877–1878. 326 + 303 pages.

Second edition. 1894–1895. 480 + 491 pages.

Reprint of the 2nd ed. London: Macmillan, 1926–1929. [Photorepr. New York: Dover, 1945.]

German translation. Die Theorie des Schalles (trans. F. Neesen), 2 vols., Braunschweig: Vieweg, 1879–1880.

Related articles: Thomson and Tait (§40), Maxwell (§44).

1 RAYLEIGH'S EARLY RESEARCH ON SOUND

Lord Rayleigh (John William Strutt, 1842–1919) was one of the most influential British scientists in the late 19th and early 20th centuries and the Nobel Laureate of Physics in 1904. He graduated from Cambridge University as senior wrangler of Mathematical Tripos in 1865 and decided to become a scientist regardless of his noble birth. Almost all his researches were performed at his manor in Terling Place, Essex, except during his professorship at Cavendish Laboratory in Cambridge from 1880 to 1884. The range of his interests covered almost all the subjects of physics including sound, light, electricity, heat, and gas. His most famous contributions to physics involved the experimental determination of electrical unit Ohm (during his time as Professor), the formulation of the Rayleigh–Jeans

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

law for black body radiation, and the discovery of the inert gas argon, which was honored by the Nobel Prize.

During all of his scientific research career, Rayleigh remained very interested in acoustical phenomena. He began to investigate acoustical oscillations experimentally and mathematically in the 1860s, just after he graduated from Cambridge University as Senior Wrangler. His interest in acoustics was stimulated by the encouragement of W.F. Donkin (1814–1869), the Savilian Professor of Astronomy at Oxford University. After reading the *Tonempfindungen* (‘*Sensations of tone*’) of Hermann von Helmholtz (1821–1894), he started experiments with Helmholtz’s resonators [Lindsay, 1945].

Strutt’s early acoustical research soon included mathematical developments, culminating in the paper ‘On the theory of resonance’. In this paper, of 1870, he introduced conductivity c, the inverse of hydrodynamical resistance, in analogy with electrical conductivity, in order to treat the vibrations of air in various tubes. In the same paper he made use of the notion of velocity potential, which had never been employed, even by Helmholtz, in treating the problem of resonance. His mathematical treatments were tested by his own and other researchers’ experimental results in the latter part of this paper. The balance between theory and experiment was to become one of the main characteristics of his acoustical research [Ku, 2002]. This paper received favorable acceptance from physicists including J.C. Maxwell (1831–1879), and his acoustical research was promoted.

In addition to his book Strutt published various papers, which are reprinted in [Rayleigh, *Papers*]. His masterly paper ‘Some general theorems relating to vibrations’, published in 1873, included three key innovations in vibration theory. First, he proposed an effective and practical approximation method to find frequencies of various systems of vibration, which was to be called the ‘Rayleigh–Ritz’ method. Second, he introduced the dissipation function, which represented the dissipation of energy in a system subject to resistant forces varying with velocity: it was later recognized as an innovation to treat systems of vibration, and applied to the analysis of electrical circuits by Maxwell [Hong, 1994, 28, 345–347]. Third, Strutt pursued the extension of the ‘reciprocal theorem’ to generalized vibratory systems. The reciprocal theorem suggests that when a system includes two parts A and B, the vibratory force acted at A produces at B the same vibration as would have ensued at A had the force acted at B. In 1860, Helmholtz had proved the reciprocal theorem only in those systems vibrating in a non-resistant homogeneous fluid. Strutt deduced that it could be extended to the cases of vibrating systems such as strings, membranes, and tuning forks in a resistant medium.

Strutt’s early acoustical investigations were characterized by generalized mathematical analyses and their experimental proofs. His paper ‘On waves’ of 1876 revealed his general interest in vibrations and waves. In this paper, he was concerned with water waves, but his mathematical treatment of some problems was related to other types of vibrations and waves. He thought that hydraulics, acoustics, optics, electricity and magnetism were intimately related to each other with regard to vibrations and waves.

2 THE PUBLICATION OF *THE THEORY OF SOUND*

While laying a firm foundation for his acoustical research in this way, Strutt began to write *The theory of sound* (hereafter, ‘*TS*’) in December 1872 in the cabin of a boat on the Nile,

where he was taking a rest cure after an attack of rheumatic fever that had nearly killed him. His excellent mathematical skills, which were imparted by his coach E.J. Routh, were highlighted in analyses of mechanical vibrations relating to sound. He aimed at a generalized mathematical treatise, which would be abreast of William Thomson's and P.G. Tait's *Treatise on natural philosophy* (1867) (§40) and Maxwell's *Treatise on electricity and magnetism* (1873) (§44).

When Strutt came back to London in 1873, it was known that he was writing a treatise on sound, and this attracted high expectations. (At his return, his father passed away and he succeeded as Third Baron Rayleigh.) Maxwell, for example, told Rayleigh in a letter of May 1873 that he expected Rayleigh's treatise would fill the gap in the paucity of English writings relating to acoustics. After proofreading by H.M. Taylor, Rayleigh's friend and competitor in the Mathematical Tripos, the first volume of *TS* was published in 1877 and the second in 1878 by Macmillan.

The academic response to this treatise was beyond expectation. Routh remarked that it was a wanted book, which he would use as a textbook; he expected that he would learn a lot from it and that it would contribute to the progress in this area. G.B. Airy, who was investigating acoustics and hydrodynamics, said that this treatise not only deeply discussed sound but also dealt with many non-acoustical vibrations, and was therefore applicable to much more complicated subjects. Above all, Helmholtz, one of the ultimate authorities on acoustics at that time, reviewed it favorably in *Nature* [Helmholtz, 1878]. He observed that this book put forth subjects in a coherent and accessible form, and thus would help acoustical research a great deal, and the methods employed were capable of promoting further progress in research in this field. He wrote that since it did not include chapters dealing with the theory of reed pipes, including the human voice and the mathematical explanation of singing flames, the blowing of organ pipes, and maintained vibrations such as the action of the violin bow and the Aeolian harp, it should not be considered complete in these two volumes.

Rayleigh admitted that a third volume was needed, and the publishers urged him to write it, but this was not realized. Instead, a considerable part of what Helmholtz pointed out was included in the revised and enlarged edition, which was published in 1894 and 1895. Helmholtz's favorable review allowed the book to be translated swiftly into German: at his suggestion, F. Neesen translated it in 1878 [Strutt, 1924].

3 THE BOOK AS COMPARED WITH ITS PREDECESSORS

The contents of Rayleigh's book are summarized in Table 1. Its main purpose was to gather mathematical investigations of sound conducted up to that time. He said in the preface that he wrote the treatise in order to correct the situation in which many research papers on sound were not comprehensive. Having judged that researchers' inability to access research materials was blocking the development of this area, he read and analyzed treatises and papers in periodicals and transactions of academies published in Great Britain, Germany, France, Switzerland, etc. *TS* was intended to be a comprehensive mathematical treatise on sound, and so complete the task which Donkin's posthumous book *Acoustics* (1870) had not fulfilled because of his sudden death in 1869. Before this work, some acoustical texts

Table 1. Summary by chapters of Rayleigh's book. Volume 2 starts at chapter XI.

Chapter	Page	'Title' or Description
I	1	Introduction.
II	18	Harmonic motions.
III	35	Systems with one degree of freedom.
IV	67	Vibrating systems in general. Generalized co-ordinates, Lagrange's equations.
V	97	Vibrating systems in general. Various kinds of forces and systems of vibration.
VI	127	Transverse vibrations of strings.
VII	188	Longitudinal and torsional vibrations of bars.
VIII	201	Lateral vibrations of bars.
IX	250	Vibrations of membranes.
X	293	Vibrations of plates. [End 326.]
XI	1	Aerial vibrations.
XII	44	Vibrations in tubes.
XIII	65	Aerial vibrations in a rectangular chamber.
XIV	85	Arbitrary initial disturbance in an unlimited atmosphere.
XV	135	Secondary waves due to a variation in the medium.
XVI	156	Theory of resonators.
XVII	204	Applications of Laplace's functions to acoustical problems.
XVIII	253	Problem of a spherical layer of air.
XIX	280	Fluid friction. [End 303.]

had been published, such as E.E.F. Chladni's *Die Akustik* (1802), volume 1 of Thomas Young's *A course of lectures on natural philosophy and the mechanical art* (1807), E.H. and Wilhelm Weber's *Wellenlehre* of 1825, Benjamin Peirce's *An elementary treatise on sound* of 1836, Helmholtz's *Tonempfindungen* [Helmholtz, 1878], and John Tyndall's *On sound* of 1867. But these acoustical texts all focused themselves on the description of empirical and experimental findings. In these acoustical texts, mathematical inquiries, which had been made by mathematicians from various countries, had been treated as subordinate. John Herschel's article 'Sound' for the *Encyclopaedia metropolitana* in 1830, though considerably mathematical, was neither comprehensive nor systematic.

Rayleigh's book fulfilled these aims and did more. Much interested in describing the history of theoretical investigations on sound by mathematicians, he tried to gather as many mathematical works on sound as possible and arranged them systematically in the treatise. His great mathematical ability enabled him to understand these mathematical investigations and to arrange them in a systematic order. Consequently, a considerable part of the theoretical investigation in *TS* was not the result of the author's own original research.

In addition, the discussion in *TS* was not limited to mathematical analyses, since Rayleigh did not want to conduct mathematical discussion in isolation of experimental achievements. Thus he collected and presented extensive materials related to experimental or empirical researches on sound in the treatise. In fact, he had embarked on his acoustical

research by performing experiments, and his experimental research continued until the end of his life. His character as an experimentalist made his book different from other contemporary mathematical treatises like Donkin's *Acoustics*. When he introduced experimental achievements, he usually described the processes in detail so that other researchers might reproduce the results. This feature was exceptional in a scientific treatise in which mathematical analyses were primary. For example, discussing preceding studies on the vibration of membranes in Chapter 9 [Ku, 2002], Rayleigh described in detail how M.J. Bourget, a pioneer in experimenting with the vibration of membranes, made an effective paper membrane and did experiment with it (art. 213):

> The paper is immersed in water, and after removal of the superfluous moisture by blotting paper is placed upon a frame of wood whose edges have been previously coated with glue. The contraction of the paper in drying produces the necessary tension, but many failures may be met with before a satisfactory result is obtained [...] If the vibration be sufficiently vigorous, the sand accumulates on the nodal lines, whose form is thus defined with more or less precision. Any inequality in the tension shews itself by the circles becoming elliptic.

TS included all kinds of acoustical information, whether theoretical or experimental. However, Rayleigh was not satisfied with simply compiling theoretical and experimental information. He instead sought to make a close connection between theoretical analyses and experimental facts, valuing a match between them as a good confirmation of his mathematical analysis. In many cases, he could not conduct exact mathematical analyses and had to take some approximation. In such cases, he justified his approximation by providing empirical and experimental proofs. But when empirical facts or mathematical analyses did not have their counterparts, he was satisfied with simply putting them together without unifying them.

4 ON RAYLEIGH'S MATHEMATICAL METHODS IN THE BOOK

One noteworthy feature of *TS* lies in his method of analysis. Rayleigh's typical method to solve problems was to formulate a differential equation for a given phenomenon. He was trained in this method through the Mathematical Tripos. He often introduced ideal conditions and added assumptions or simplified models in order to find a differential equation describing a system of vibration or wave. Nevertheless, there were many cases in which solving the equations was difficult because of the limitation of mathematics. In order to overcome such difficulties, one of his most powerful and frequent strategies was the method of successive approximation.

In Chapter 3, for example, after solving the one-dimensional vibration of an acoustical system which was restricted to small displacements, Rayleigh extended his discussion to a general case where the second- and higher-order terms had to be considered. He put kinetic and potential energies as

$$T = \frac{1}{2}(m_0 + m_1 u)\dot{u}^2 \quad \text{and} \quad V = \frac{1}{2}(\mu_0 + \mu_1 u)u^2, \tag{1}$$

where u is the displacement. The equation of motion was obtained by differentiating the sum of T and V,

$$m_0\ddot{u} + \mu_0 u + m_1 u\ddot{u} + \frac{1}{2}m_1\dot{u}^2 + \frac{3}{2}\mu_1 u^2 = \text{Impressed Force.} \tag{2}$$

As this could not be easily solved in this form, he let $m_1 = 0$, so that the mass was independent of the displacement u. The equation was reduced to

$$\ddot{u} + n^2 u + \alpha u^2 = 0. \tag{3}$$

Then an approximate solution obtained by neglecting the last term of the left member,

$$u = A\cos nt \tag{4}$$

was substituted for the last term neglected in the original equation, and he arrived at the equation

$$\ddot{u} + n^2 u = -\alpha\frac{A^2}{2}(1 + \cos 2nt). \tag{5}$$

From this he obtained a modified approximate solution

$$u = A\cos nt - \frac{\alpha A^2}{2n^2} + \frac{\alpha A^2}{6n^2}\cos 2nt, \tag{6}$$

which implied that the system produced the vibrations not only of the fundamental angular frequency n, but also of the first harmonic angular frequency $2n$. Rayleigh justified this process by pointing out that the higher frequency could be perceived when a tuning fork was hit violently.

5 RAYLEIGH ON WAVES AND VIBRATIONS

Another outstanding feature of *TS* was its great concern with the general theories of waves and vibrations. Although Rayleigh's principal purpose for writing this treatise was to present completely the theoretical analyses of acoustical phenomena, *TS* included non-acoustical discussions, since his concern over waves and vibrations was broad. He presented generalized theories applicable to optical phenomena, electrical vibrations, tides, water waves, and perturbations of celestial bodies as well as acoustical phenomena. He preferred generalized discussion, because not only they were applicable to various problems but they also revealed the unity of nature.

Thus, in dealing with acoustical vibrations, Rayleigh frequently discussed other phenomena related to vibrations as well. While examining the difference in motions caused by violent forces, for instance, he exemplified an optical extension of acoustical phenomena by showing the selective absorption of the two kinds of light. Again, in pressing a close analogy between optical and acoustical phenomena over the reflection and the refraction of waves in Chapter 13, he derived the formula of A.J. Fresnel for light polarized perpendicular to the plane of incidence, and obtained the conditions for the total reflection obtained

by Sir David Brewster. Then he ascertained the general conditions for the reflection and the refraction of waves by mathematical analysis. The result was a general theory which was applicable to light as well as sound.

In the center of the general treatments were differential equations that were repeatedly used in the text. For Rayleigh, some differential equations applicable to several problems linked the corresponding phenomena, and paved the way to general theories. Different phenomena, which had been regarded as distinct in experiments, could be considered as closely related owing to common equations of motion. This relation had not been perceived in experimental acoustics. The mathematical discussion enabled him to understand the hidden order behind the phenomena. One of the equations that were employed several times and connected important phenomena was the basic wave equation

$$\frac{d^2y}{dt^2} = a^2\frac{d^2y}{dx^2}. \tag{7}$$

Rayleigh first used this equation when he treated the propagation of the transversal wave in an infinitely extended string. Yet the same form was found in the description of both the longitudinal and the torsional vibrations of the rod. These vibrations were caused by entirely different forces and properties of the media. In this manner, the experimental results of entirely different areas could be explained on the same mathematical basis.

Another differential equation, which appeared often in volume 2, was Laplace's equation (compare §18.5). In Chapter 11, Rayleigh derived the equation with regard to the velocity potential in discussing aerial vibration generally. In this treatise, this equation was widely used in treating the transmission of sound in various spaces. Bessel's equation was also widely employed in this treatise in describing several vibrations related to the circular membrane or circular plate. Their specific solutions, which were already widely known, described acoustical and vibrational phenomena in the same way as they did other phenomena. In this fashion, the use of these versatile differential equations helped Rayleigh to develop general theories.

6 PRESENTING ORIGINAL RESEARCHES

A considerable portion of *TS* was devoted to presenting Rayleigh's original theoretical or experimental investigations. His originality was most remarkable in Chapters 4 and 5, in which he treated the general theory of vibration by introducing generalized coordinates. In Chapter 4 he deduced Lagrange's equation by using generalized coordinates and by employing d'Alembert's principle (§11.3) and the principle of virtual velocities (compare §16.3 on Lagrange). He obtained

$$\frac{d}{dt}\left(\frac{dT}{d\dot{\psi}}\right) - \frac{dT}{d\psi} = \Psi, \tag{8}$$

where T was kinetic energy, ψ a generalized coordinate, and Ψ the generalized component of force. In the case of a conservative system, by separating from Ψ those parts depending

only on the configuration of the system, he changed the equation into

$$\frac{d}{dt}\left(\frac{dT}{d\dot{\psi}}\right) - \frac{dT}{d\psi} + \frac{dV}{d\psi} = \Psi, \tag{9}$$

where V was the potential energy of the conservative system and Ψ was confined only to the force which was not derived from potential energy. Rayleigh rewrote the equation in the form

$$\frac{d}{dt}\left(\frac{dT}{d\dot{\psi}}\right) - \frac{dT}{d\psi} + \frac{dF}{d\dot{\psi}} + \frac{dV}{d\psi} = \Psi \tag{10}$$

which included the dissipation function F, representing the effect which was produced by friction or viscosity. From the linearity of this equation, he derived Daniel Bernoulli's principle of the coexistence of small motions, according to which the second order terms could be neglected when small vibrations are superposed. Rayleigh showed that this was applicable to problems of one-dimensional vibration by adding constraints to generalized problems.

Rayleigh's original investigations continued in Chapter 5. He treated the cases in which F was so simple that the general equation of motion could be reduced to the form as for a system of one degree of freedom. On this condition, he obtained an equation of motion in the following form

$$a\ddot{\phi} + b\dot{\phi} + c\phi = \Phi, \tag{11}$$

where a, b, c are arbitrary constants and ϕ velocity potential. From this he obtained solutions for damped free vibrations by using the condition of $\Phi = 0$, solutions for violent vibration depending on Φ, and simpler solutions for violent vibrations without friction. He considered the practical example of a stretched string with a harmonic force acting on a point of it. In the next part of the chapter, Rayleigh deduced the general reciprocal theorem from the properties of the corresponding partial differentials and determinants, and applied it to specific examples. He proved that the reciprocal theorem was applicable to the case of the existence of the dissipation function F. He pointed out that the theorem was not applicable in such cases as the transmission of sound waves during the wind blowing, for the theorem was applicable only to the vibration around the arrangements of equilibrium.

It was in Chapters 14 to 17 that Rayleigh's discussions were most original. In Chapter 14, he introduced his own experiment in discussing sound transmission, in order to test the theory of the interference of sound waves. In the experiment, two 256-Hertz tuning forks 10 yards distant from each other were driven by an intermittent electric current made by a 128-Hertz tuning fork interrupter. This experiment was first reported in *Philosophical magazine*, published in June 1877. After strengthening the intensities of the sound by the addition of resonators to tuning forks, he detected the points of silence at places where the theory indicated, which confirmed his own expectation.

In Chapter 15, Rayleigh put forth his original theory on the secondary waves that are produced when the plane waves impinge on different media. He deduced that the amplitude of the secondary waves varies inversely to the distance through the medium and to the square value of the wavelength, and that while a region in which the compressibility varies acts like a simple source, a region at which the density varies acts like a double source.

He illustrated these theoretical arguments by means of harmonic echoes, on which he had written a paper in 1873.

In Chapter 16, on the theory of resonators, Rayleigh presented the results of mathematical and experimental investigations on resonators he had performed. A considerable portion of the chapter came from his paper 'On the theory of resonance' of 1870.

Chapter 17 included discussion on sound waves that are generated as a reaction to the vibration of a rigid body and propagate in the air. Referring to George Green (§30.5) and S.D. Poisson as predecessors, Rayleigh developed his own analytical theory. He solved Laplace's equations in various situations, employing such functions as spherical harmonics, Legendre's functions, and Bessel's functions. He proceeded from this to the case of disturbance confined to a small portion of a spherical surface. By using the reciprocal theorem, he transformed this problem into that of a sound wave which arrived at any point on the spherical obstacle from an external source, and he could therefore discuss the obstructive effect of a head in the path of the transmission of sound in the air. Then, Rayleigh discussed the application of the general equations when there was no sound source. He employed the theory of Bessel functions for the problem of no source at the pole, and applied the result successfully to the vibration of the gas in a spherical rigid envelope.

In this way *TS* included almost all of Rayleigh's original achievements in acoustics up to then. Together with other researchers' work, it constituted a comprehensive system of acoustics.

7 THE INFLUENCE OF THE BOOK ON ACOUSTICS AND ELSEWHERE

When the first American edition of *TS* was published in 1945, American acoustician R.B. Lindsay said in the 'Historical introduction' that the treatise was being used as a standard text in the acoustical arena, though it was first published more than 65 years earlier [Lindsay, 1945]. *TS* contains such a useful contents for acousticians that, as late as the 1980s, acoustician Thomas Rossing said: 'I do not know of a musical acoustician who does not keep a well-thumbed copy of [Rayleigh's book] in his/her personal library'. For R.T. Beyer 'One can rarely pick up either the book or the collected papers without finding something of real interest that was previously missed' [Beyer, 1999, ch. 4]. This enduring influence of Rayleigh's book began to take shape shortly after the first publication in 1877. What made *TS* so special to the acousticians?

First of all, the information included in *TS* was remarkably varied and abundant. Acousticians found subjects of their researches in *TS* and could contribute by adding new elements to Rayleigh's ideas. They also found invaluable information and materials which could not be found in other acoustical writings. As Lindsay [1945] said, it was a 'mine of information'. For example, in 1886, S.P. Thompson pointed out the superiority of Rayleigh's arrangement of the electromagnetically-driven tuning fork as described in *TS*. His arrangement with a short magnetic rod placed between the prongs was different from that of earlier apparatuses, with a U-shaped magnet outside the prongs of the tuning fork. According to Thompson, Rayleigh's was a more effective method by which energy was transmitted to the fork. Soon this arrangement was adopted by other acousticians in their own apparatus. And, in 1882, American acoustician John LeConte interpreted his sound

shadow experiment in water by following Rayleigh's interpretation. In 1884, D.J. Blaikley quoted theoretical results in *TS* in reporting the experiment on the velocity of sound in the tube and considered Rayleigh an important authority. Before the turn of the century, *TS* had become a standard text for acousticians.

In addition, mathematical methods introduced in *TS* became guides for theoretical acoustical research in the late 19th and early 20th century. In 1907, for example, by extending Rayleigh's mathematical treatment of the scattering of sound by a sphere, J.W. Nicholson treated the scattering of sound by a spheroid and a disk. In 1908, E.H. Barton discussed the propagation of sound in conical pipes by referring to the contents of *TS* and taking such methods presented in *TS* as method of dimensions and successive approximation. This proves that *TS* had become an exemplar in the area of acoustical research. Many acousticians took *TS* as a starting point of their investigations and no other book could play the role in this area at that time.

The impact of *TS* on the German-speaking scientific community was as strong as on the English-speaking one. German acousticians could easily refer to the German translation, *Die Theorie des Schalles*. In the late 19th and early 20th centuries, many German acousticians such as Heinrich Kayser, A. Oberbeck, A. Elsas, Max Wien, Georg Stern and P. Drude viewed *Die Theorie des Schalles* as a main authority or found a foundation and information for their research in it.

In this process, *TS* created a unified image of acoustics in the minds of acoustical researchers and physicists. Before the publication of *TS*, acoustical researchers had been more or less separated in two camps, that is to say, experimentalists and mathematicians. The experimental tradition was shaped from Chladni's famous figures and Young's acoustical discussion in the early 19th century. They gave an impression upon researchers that 'acoustics' was an area of experimental research. Many 'acousticians' were engaged in empirically gathering 'facts' related to sound. On the other hand, since Newton calculated the velocity of sound, acoustical phenomena became the concerns of mathematicians. During the 18th and the 19th centuries, mathematicians like d'Alembert, Leonhard Euler, Daniel Bernoulli, Lagrange, Poisson, Sophie Germain, Gustav Kirchhoff, and G.G. Stokes attempted to analyze various vibrating bodies mathematically [Beyer, 1999]. But almost all acoustical texts published before *TS* omitted more or less these mathematical achievements. Most experimentalists were little concerned with mathematical analyses. Helmholtz, an acoustical experimentalist and mathematician at once, was exceptional, but even his *Tonemfindungen* relegated mathematical analyses to appendices [Vogel, 1993].

The experimental and the mathematical traditions were connected and became complementary in *TS*. Afterwards acoustical researchers followed the style of research manifested there, considering theoretical investigations as closely related to empirical findings. They became interested in both sides, though some of them did research in only one side. (For example, the pioneering works on architectural acoustics of the American acoustician Wallace Sabine in the early 20th century were mainly experimental but reflected both traditions.) Whether experimentalists or mathematicians, acoustical researchers consulted *TS*, and began to recognize that they were working in one field of 'acoustics'.

The influence of *TS* went beyond acoustics and reached general physics and engineering [Humphrey, 1992]. For Rayleigh created some new mathematical techniques, and his derivations and developments were used in other fields of physics and engineering later

[Beyer, 1999, chs. 7–10]. When the reprint of the second edition was issued in 1926, Harvey Fletcher of Bell Telephone Laboratories pointed out that this book was being used by researchers concerned with electric vibrations, especially telephone engineers [Fletcher, 1928]. It was because *TS* included comprehensive problems related to vibrations, not restricting its discussion to audible acoustical vibrations as its title suggested. Rayleigh's general theories expanded in the second edition, including electric vibrations (Chapter 10B), capillarity (Chapter 20), vortex motion and sensitive jets (Chapter 21). In the new chapters, equations and concepts which were created in order to express acoustical vibrations were applied to electrical transmissions, water waves, vortex motion, sensitive jets, and so forth.

One of the non-acoustical areas in which the impact of this treatise was the most noteworthy was the theory of elasticity. Rayleigh's mathematical treatment of the motions of various sonorous objects became a foundation for analyses of vibrations of solid elastic bodies. For example, the influence of *TS* on A.E.H. Love's *Treatise on the mathematical theory of elasticity* of 1892–1893, which was to become a classic of the theory of elasticity during the 20th century, was enormous. Rayleigh's mathematical analyses of vibrations of various forms of bodies such as rods, circular plates, cylinders, and curved plates were employed as an essential basis for Love's discussion.

Rayleigh's pursuit of generality was one of the main causes of his influence on various fields. His treatments of acoustical systems in *TS* aimed at general vibrations and waves, and they could be easily applied to other problems. Various methods and notions in *TS* have become fundamental in physics and other related fields. The Rayleigh–Ritz method, which is widely used for approximation in quantum mechanics, and the notion of state density, which is now an essential concept in solid-state physics, could be considered as representative examples.

BIBLIOGRAPHY

Beyer, R.T. 1999. *Sounds of our times: two hundred years of acoustics*, New York: Springer.

Fletcher, H. 1928. Review of Rayleigh's book, 2nd ed. reprint, *Proceedings of the Institute of Radio Engineers, 16*, 181–191.

Hong, S. 1994. 'Forging the scientist–engineer: A professional career of John Ambrose Fleming', unpublished Ph.D. Dissertation, Seoul National University.

Helmholtz, H. 1878. Review of Rayleigh's book, 1st ed., *Nature, 17*, 237–239; *19*, 117–118. [Repr. in [Beyer, 1999], 419–428.]

Humphrey, A.T. 1992. 'Lord Rayleigh—the last of the great Victorian polymaths', *GEC review, 7*, 167–179.

Ku, J.H. 2002. 'Rayleigh (1842–1919) ui umhyanghak yeonku-ui seongkyeok-gwa seongkwa' ['The characteristics and accomplishment of Lord Rayleigh's acoustical research'], unpublished Ph.D. Dissertation, Seoul National University.

Lindsay, R.B. 1945. 'Historical introduction', in Rayleigh, *The theory of sound*, repr. New York: Dover.

Lindsay, R.B. 1970. *Lord Rayleigh, the man and his works*, London: Pergamon Press.

Lindsay, R.B. 1976. 'Strutt, John William, Third Baron Rayleigh', in C.C. Gillispie (ed.), *Dictionary of scientific biography*, New York: Scribners, vol. 13, 100–107.

Miller, D.C. 1935. *Anecdotal history of the science of sound: to the beginning of the twentieth century*, New York: Macmillan.

Rayleigh, Lord. *Papers. Scientific papers*, 6 vols., Cambridge: Cambridge University Press. [Repr. New York: Dover, 1964.]

Strutt, R.J. 1924. *Life of John William Strutt, Third Baron Rayleigh, O.M., F.R.S.*, London: Edward Arnold.

Vogel, S. 1993. 'Sensation of tone, perception of sound, and empiricism: Helmholtz's physiological acoustics', in D. Cahan (ed.), *Hermann von Helmholtz and the foundations of nineteenth-century science*, Berkeley: University of California Press, 259–287.

CHAPTER 46

GEORG CANTOR, PAPER ON THE 'FOUNDATIONS OF A GENERAL SET THEORY' (1883)

Joseph W. Dauben

In this revolutionary monograph, Georg Cantor set out the earliest detailed version of his transfinite set theory, including a theory of transfinite ordinal numbers and their arithmetic; and a defense of the theory on historical and philosophical grounds. In concert with his later articles on the foundations of set theory (1895–1897) it created a virtually new discipline, set theory.

First publication. *Grundlagen einer allgemeinen Mannigfaltigkeitslehre. Ein mathematisch-philosophischer Versuch in der Lehre des Unendlichen*, Leipzig: Teubner, 1883. 46 pages. Also publ. (without the preface) as 'Über unendliche, lineare Punktmannichfaltigkeiten', part 5, *Mathematische Annalen, 23* (1884), 453–488.

Reprint. In Cantor, *Gesammelte Abhandlungen mathematischen und philosophischen Inhalts* (ed. E. Zermelo), Berlin: J. Springer, 1932 (repr. Hildesheim: Olms, 1966; Berlin: Springer, 1990), 165–208 [cited below].

Partial French translation. 'Fondements d'une théorie générale des ensembles', *Acta mathematica, 2* (1883), 381–408.

English translations. 1) 'Foundations of the theory of manifolds' (trans. U. Parpart), *The campaigner* (The Theoretical Journal of the National Caucus of Labor Committees), *9* (January and February 1976), 69–96. 2) and preferable: 'Foundations of a general theory of manifolds: a mathematico-philosophical investigation into the theory of the infinite', by W.B. Ewald in his (ed.), *From Kant to Hilbert: a source book in the foundations of mathematics*, New York: Oxford University Press, 1996, vol. 2, 878–920.

Russian translation by P.S. Yushkevich in F.A. Medvedev and others (ed.), Cantor, *Trudi po teorii mnozhestv*, Moscow: Nauka, 1985, 63–106.

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

Italian translation by G. Rigamonti in (ed.), Cantor, *La formazione della teoria degli insiemi* (*saggi 1872–1883*), Florence: Sansoni, 1992, 77–127.

Related articles: Cauchy and Riemann on real-variable analysis (§25, §38), Dedekind (§43, §47), Hilbert on problems (§57), Whitehead and Russell (§61), Lebesgue and Baire (§59).

1 CANTOR'S WAY IN

Georg Ferdinand Ludwig Philip Cantor was born on 3 March 1845 in St. Petersburg, Russia. His mother, a Roman Catholic, came from a family of notable musicians; his father, the son of a Jewish businessman from Copenhagen, was also a successful tradesman with commercial interests extending as far as Latin America. Although the exact circumstances are unknown, Cantor's father was raised in a Lutheran mission in St. Petersburg, and he passed his own deeply religious beliefs on to his son. Later in life, Cantor's religious convictions would play a significant role in his steadfast faith in the correctness of his controversial transfinite set theory, just as his mother's Catholicism may have made him particularly amenable to the substantial correspondence he undertook with Catholic theologians over the nature of the infinite from a theological perspective. On his life and early career see [Dauben, 1979, especially pp. 271–299]; other biographical studies include [Meschkowski, 1967; Grattan-Guinness, 1971], and [Purkert and Ilgauds, 1985, 1987].

The family moved from Russia to Germany when Cantor was a boy; it was there that he went to university, where he studied mathematics in Berlin, and briefly at Göttingen. After receiving his doctorate from the University of Berlin in 1868 for a dissertation on the theory of numbers, two years later he accepted a position as *Privatdocent*, or instructor, at the University of Halle. Among his colleagues there was Eduard Heinrich Heine (1821–1881), who was then working on the theory of trigonometric series. Given an arbitrary function represented by a trigonometric series, the question of the uniqueness of the representation was a difficult challenge (essentially set by Bernhard Riemann: §38.4), and Heine encouraged Cantor to consider this problem. Heine offered a partial solution in 1870 for the case of almost everywhere continuous functions, assuming as well the uniform convergence of the trigonometric series in question. Cantor sought to do away with these restrictions, and to establish the uniqueness theorem on the most general terms possible.

2 EARLY WORK ON TRIGONOMETRIC SERIES: DERIVED SETS

Cantor at first succeeded in proving the uniqueness theorem for an arbitrary function

$$f(x) = \sum_n (a_n \sin nx + b_n \cos nx) \tag{1}$$

in the case that the trigonometric series was convergent for all values of x. But in 1871 he went further, and published a note showing that the uniqueness theorem held even if, for certain values of x, either the representation of the function or the convergence of the trigonometric series were given up, so long as the number of such exceptional points x was finite.

Cantor's major publication on these matters came a year later—a substantial paper showing that the uniqueness theorem held even in the case of an infinite number of exceptional points, so long as such points were distributed in a specified way [Cantor, 1872]. It was the exact determination and analysis of this specification that opened the door not only to his development of the theory of point sets, but to his later theory of the transfinite ordinal and cardinal numbers as well [Dauben, 1971].

The specification that Cantor introduced in his paper of 1872 concerned the limit points of an infinite set P. Given any such set, by the Bolzano–Weierstrass theorem it must contain at least one point, any arbitrarily small neighborhood of which contains an infinite number of points. The set of all such limit points of P Cantor denoted P', the first derived set of P. If P' also contains an infinite number of points, it too must contain at least one limit point, and the set of all its limit points P'' is the second derived set of P. Continuing his consideration of derived sets, if for some finite number ν the νth derived set P^{ν} is not an infinite set, then its derived set, the $(\nu+1)$th derived set of P, $P^{\nu+1}$, will be empty, i.e. $P^{\nu+1} = \emptyset$. Cantor called such sets derived 'sets of the first species', and for such sets of exceptional points of the first species, he was able to show that his uniqueness theorem for trigonometric series representations remained valid. As yet, he did not know what to make of derived sets of the second species, but these would soon begin to attract his attention, with remarkable and unexpected consequences.

3 CANTOR'S THEORY OF REAL NUMBERS

Meanwhile, Cantor's introduction of point sets of the first species required that he be able to specify their structure in a definite way. This in turn meant that he needed a rigorous concept of the real numbers in general, a subject he also considered in his paper [1872] on trigonometric series. There he introduced real numbers in terms of convergent sequences of rational numbers, and realized that he had to take as axiomatic that there was a one-to-one correlation between the real numbers and the points on the line, namely that the arithmetic and geometric continuums are comparable. How to account for the continuum of real numbers was a problem that would continue to haunt Cantor for the rest of his career, especially in terms of his famous conjecture, the Continuum Hypothesis, namely that the infinite set of real numbers R was the next higher order of infinite sets after denumerably infinite sets like the set of all natural numbers N. But in 1872 Cantor was not at a point yet where he could even formulate this proposition in any meaningful, precise way.

Cantor was not alone in studying the properties of the continuum of real numbers in rigorous detail. In 1872, the same year in which his paper appeared, the German mathematician Richard Dedekind (1831–1916) also published an analysis of the continuum that was based on infinite sets (§43). Dedekind articulated an idea that Cantor later made more precise: that 'the line L is infinitely richer in point-individuals than is the domain R of rational numbers in number-individuals' [1872, 9].

But neither Dedekind nor Cantor was in a position to say how much richer the infinite set of points in the continuum was than the infinite set of rational numbers. Cantor's next contribution to this question was published as [Cantor, 1874], in Crelle's *Journal für die reine und angewandte Mathematik*: five pages on the non-denumerability of the real numbers.

Surprisingly, in its title 'On a property of the collection of all real algebraic numbers' Cantor mentioned only the algebraic numbers, not the set of all real numbers. Nevertheless, in this paper he disclosed his revolutionary discovery of the non-denumerability of the continuum of real numbers. This meant that in an absolute sense some infinite sets were larger than others, in particular that the set of natural numbers N was of a lower magnitude of infinity than the set of real numbers R. Cantor's reasons for not mentioning this result in the title of his paper may have been due to his fear that any suggestion that the real numbers were non-denumerably infinite would prove extremely controversial. But what in particular could have prompted him to choose such an inappropriate title, concealing one of his most remarkable discoveries, one that in retrospect strikes any mathematician as among the most important discoveries in modern mathematics?

The answer hinges on one of Cantor's teachers at Berlin, Leopold Kronecker (1823–1891), who also edited Crelle's journal. Having studied with Kronecker, Cantor was well acquainted with his work in number theory and algebra, and with his highly conservative philosophical views with respect to mathematics. By the early 1870s, Kronecker was already vocal in his opposition to any infinitary arguments, including the Bolzano–Weierstrass theorem, upper and lower limits, and to irrational numbers in general. Kronecker's later pronouncements against analysis and set theory, as well as his adamant insistence upon using the integers to provide the only satisfactory foundation for mathematics, were simply extensions of these early views [Edwards, 1989]. It is not unreasonable to suspect that Cantor had good reason to anticipate Kronecker's opposition to his proof of the non-denumerability of the real numbers, which proved they comprised a set infinitely larger than the set of integers [Dauben, 2005].

There was a positive side, however, to Kronecker's opposition to Cantor's work; it forced Cantor to evaluate the foundations of set theory as he was in the process of creating it. This concern prompted long philosophical passages in Cantor's major publication of the 1880s on set theory, our landmark: his *Grundlagen einer allgemeinen Mannigfaltigkeitslehre* of 1883. It was there that Cantor issued one of his most famous pronouncements about mathematics, namely that 'the *essence* of *mathematics* lies precisely in its *freedom*' ([Cantor, 1883, 182]: unfortunately the frequent quotations of this motto usually lack the important word 'precisely'). This was not simply an academic or philosophical message to his colleagues, for it carried as well a hidden and deeply personal subtext. It was, as he later admitted to David Hilbert (1862–1943), a plea for objectivity and openness among mathematicians. This, he said, was directly inspired by the oppression and authoritarian closed-mindedness that he felt Kronecker represented, and worse, had wielded in a flagrant and damaging way against those he opposed.

Thus at the very beginning of his career, and even before he had begun to develop any of his more provocative ideas about transfinite set theory, Cantor was already concerned about Kronecker's opposition to his work. Doubtless he knew that more trouble could be expected in the future.

4 THE DESCRIPTIVE THEORY OF POINT SETS

Meanwhile, Cantor devoted himself to further developing his ideas about point sets that he had first investigated in the context of representing functions by trigonometric series

in the 1870s. He published the first of these in 1879, returning to his concept of derived set as a means of illuminating properties of the continuum. He defined such concepts as everywhere-dense sets, and showed that everywhere-dense sets were necessarily of the second species. Conversely, first species sets could never be everywhere-dense. But the important concept Cantor introduced in this paper was the concept of power: 'Two sets M and N are of the same power if to every element of M one element of N corresponds, and conversely, to every element of N one element of M corresponds' ([Cantor, 1879]; cited from [Cantor, *Papers*, 141]).

The two cases of greatest interest for Cantor were denumerably infinite sets of power equivalent to the set of natural numbers N, and continuous, nondenumerably infinite sets like the real numbers R. He explained the importance of the new concept of power as follows ([1879], from [Cantor, *Papers*, 150, 152]):

> The *concept of power*, which includes as a special case the concept of whole number, that foundation of the theory of number, and which ought to be considered as the most general genuine origin of sets ['*Moment bei Mannigfaltigkeiten*'], is by no means restricted to linear point sets, but can be regarded as an attribute of any *well-defined* collection, whatever may be the character of its elements [...] *Set theory* in the conception used here, if we only consider mathematics for now and forget other applications, includes the areas of arithmetic, function theory and geometry. It contains them in terms of the concept of power and brings them all together in a higher unity. *Discontinuity* and *continuity* are similarly considered from the same point of view and are thus measured with the same measure.

The following year, Cantor wrote the second paper in his series on linear point sets, which also introduced for the first time his transfinite numbers. Considering an infinite set P of the second species, it gave rise to an infinite sequence of derived sets:

$$P', P'', \ldots, P^{\nu}, \ldots . \tag{2}$$

Cantor defined the intersection of all these sets as P^{∞}. But this now led to a new sequence of derived sets, for if P^{∞} was infinite, it then gave rise to the derived set $P^{\infty+1}$. Assuming all of the subsequent derived sets were infinite, then the following sequence of derived sets was possible:

$$P', P'', \ldots, P^{\nu}, \ldots, P^{\infty}, P^{\infty+1}, \ldots . \tag{3}$$

In the paper of 1880, Cantor's focus was still on the sets themselves, and not on the 'infinite symbols' he used to specify each of the successive derived sets beginning with P^{∞}. Within months, he would begin to identify these symbols as transfinite ordinal numbers.

5 THE *GRUNDLAGEN*: A GENERAL THEORY OF SETS AND TRANSFINITE ORDINAL NUMBERS

The major achievement of the *Grundlagen* was its presentation of the transfinite ordinal numbers as a direct extension of the real numbers. Cantor admitted that his new ideas

might seem strange, even controversial, but he had reached a point in his study of the continuum where the new numbers were indispensable for further progress. Despite his own doubts at first, he said he felt forced to accept the new numbers as a legitimate and consistent part of mathematics (p. 165):

> So daring as this may seem, I can express not only the hope but the firm conviction, that this extension will, in time, have to be regarded as a thoroughly simple, appropriate, and natural one. But I in no way hide from myself the fact that with this undertaking I place myself in a certain opposition to widespread views about the mathematical infinite and to frequently advanced opinions on the nature of number.

Cantor had finally come to the realization that his 'infinite symbols' were not just indices for derived sets of the second species, but could be regarded as actual transfinite numbers that were just as real mathematically as the finite natural numbers. As he put it: 'I will define the infinite real whole numbers in the following, to which I have been led over the past few years without realizing that they were concrete numbers of real meaning' (p. 166). In order to define his new transfinite ordinal numbers independently of the derived sets of the second species, Cantor relied upon two principles of generation.

The first principle was the familiar extension of the sequence of natural numbers $1, 2, 3, \ldots$, which had its origin in the repeated addition of units. Although this sequence had no largest element, it was possible to conceive of a new number, ω, which expressed the natural, regular order of the entire sequence of natural numbers. This new number was the first transfinite number, the first number following the entire sequence of natural numbers ν. Having defined ω, it was then possible to apply the first principle of generation again to produce another sequence of transfinite ordinal numbers as follows:

$$\omega, \omega + 1, \omega + 2, \ldots, \omega + \nu, \ldots . \tag{4}$$

Again, since there was no largest element, it was possible to introduce another new number, 2ω, coming after all the numbers in the above sequence, and in fact representing the entire sequence. Cantor explained his second principle of generation, adding new numbers whenever a given sequence was limitless, as follows (p. 196):

> I call it the *second principle of generation* of real whole numbers and define them more precisely: if any definite succession of defined whole real numbers exists, for which there is no largest, then a new number is created by means of this second principle of generation which is thought of as the *limit* of those numbers, that is, it is defined as the next number larger than all of them.

By successively applying the two principles of generation, it was possible to produce an infinite sequence of numbers, the most general term being

$$\nu_0\omega^{\mu} + \nu_1\omega^{\mu-1} + \cdots + \nu_{\mu}. \tag{5}$$

To his two principles of generation, Cantor added a 'principle of limitation' meant to impose an order of sorts upon the seemingly endless hierarchy of transfinite ordinal numbers. This made it possible to identify natural breaks in the sequence, and to distinguish

between the first denumerably infinite number class of natural numbers (I), the second number-class (II), and successively higher classes of transfinite ordinal numbers (p. 197):

> We define therefore the second number-class (II) as the collection of all numbers α (increasing in definite succession) which can be formed by means of the two principles of generation:
>
> $$\omega,\ \omega+1,\ \ldots,\ \nu_0\omega^{\mu}+\nu_1\omega^{\mu-1}+\cdots+\nu_{\mu},\ \ldots,\ \omega^{\omega},\ \ldots,\ \alpha,\ldots,$$
>
> with the condition that all numbers preceding α (from 1 on) constitute a set of power equivalent to the first number class (I).

Not only did the *Grundlagen* establish that the powers of the two number classes (I) and (II) were distinct, but that (II) was the next larger after (I) (pp. 197–201).

Another important distinction that Cantor drew in the *Grundlagen* was between number (*Zahl*) and numbering (*Anzahl*). The former was simply the cardinal number or power of a given set; *Anzahl* represented both the cardinality and the order of the set in question. Whereas the two concepts coincided for finite sets, they were remarkably and significantly different for infinite sets. For example, each of the following sets of numbers were of the same cardinality or power, yet each had a distinct *Anzahl* or order:

$$(a_1, a_2, \ldots, a_n, a_{n+1}, \ldots) = \omega, \tag{6}$$

$$(a_2, a_3, \ldots, a_{n+1}, a_{n+2}, \ldots, a_1) = \omega + 1, \tag{7}$$

$$(a_3, a_4, \ldots, a_n, \ldots, a_1, a_2) = \omega + 2, \tag{8}$$

$$(a_1, a_3, a_5, \ldots, a_2, a_4, a_6, \ldots) = \omega + \omega = 2\omega. \tag{9}$$

All of the sets on the left are of the same power or cardinality—they have the same cardinal number—but by rearrangement of the same terms, each gives rise to a different *Anzahl* or ordinal number.

In the *Grundlagen*, Cantor developed an entire arithmetic for his transfinite ordinal numbers. He also discussed their arithmetic properties, above all, that they were not commutative with respect to either addition or multiplication:

$$\begin{aligned} 2+\omega &= (1, 2, a_1, a_2, \ldots, a_n, a_{n+1}, \ldots) \\ &\neq (a_1, a_2, \ldots, a_n, a_{n+1}, \ldots, 1, 2) = \omega + 2, \end{aligned} \tag{10}$$

$$2\omega = (a_1, a_2, a_3, \ldots, b_1, b_2, b_3, \ldots) \neq (a_1, b_1, a_2, b_2, a_3, b_3, \ldots) = \omega 2. \tag{11}$$

The distinction between number and numbering brought new insights to the differences between finite and infinite sets. For finite sets, the two concepts were the same. But infinite sets were more interesting because sets of the same power could have different numberings. The numbering of a set was dependent upon the order in which the elements of the set occurred. Nevertheless, there was a connection between the two: 'Every set of the power of the first class is denumerable by numbers of the second number-class, and only by such numbers' (p. 169).

These differences also explained why it was illegitimate to require infinite sets to behave exactly as did finite sets, and Cantor hoped this might help to eliminate some of the objections to the infinite in mathematics that had persisted for centuries. Indeed, it provided an alternative way of defining finite sets: if a set were finite, then its cardinal and ordinal numbers were the same (pp. 168–169).

6 THE CONTINUUM HYPOTHESIS

One goal of Cantor's transfinite set theory was to provide a means of resolving the hypothesis, that the set of real numbers R was the next in power following the set of natural numbers N. Cantor could now reformulate this conjecture more precisely, namely that the power of the continuum was equivalent to that of the second number class (II) as defined in the *Grundlagen*.

Although optimistic that the transfinite ordinal numbers and various distinctions between different kinds of point sets might soon help to resolve the Continuum Hypothesis, no solution was forthcoming. Despite Cantor's vigorous efforts to prove its correctness, he was greatly frustrated by his inability to do so. Early in 1884 he thought he had found a proof, but then reversed himself completely and thought he could actually disprove the hypothesis. Finally, he realized that he had made no progress at all, as he reported in letters that same year to his friend and editor of *Acta mathematica*, Gösta Mittag-Leffler (1846–1927) in Stockholm [Meschkowski, 1967, 240–243; Schoenflies, 1927, 12, 17–18]. All the while Cantor was facing mounting opposition and threats from Kronecker, who said he was preparing an article in which he would show that 'the results of modern function theory and set theory are of no real significance' [Schoenflies, 1927, 5].

7 CANTOR'S NERVOUS BREAKDOWNS

It was in May 1884, quite suddenly, that Cantor suffered the first of his serious nervous breakdowns that were to plague him for the rest of his life. Although his lack of progress on the Continuum Hypothesis or stress from Kronecker's attacks may have helped to trigger the breakdown, it now seems clear that such events had little to do with its underlying cause. The illness took over with startling speed and only lasted for somewhat more than a month. At the time, only the manic phase of manic-depressive psychosis was recognized as a symptom [Grattan-Guinness, 1971; Charraud, 1994]. When Cantor 'recovered' at the end of June 1884, and entered the depressive phase of his illness, he complained that he lacked the energy and interest to return to rigorous mathematical thinking. He was content to take care of trifling administrative matters at the university, but felt capable of little more.

Although Cantor eventually returned to mathematics, he also became increasingly absorbed with other interests. He undertook a study of English history and literature, and became engrossed in a scholarly diversion that was taken with remarkable seriousness by many people at the time: the supposition that Francis Bacon was the true author of Shakespeare's plays. Cantor also tried his hand without success at teaching philosophy instead of mathematics, and he began to correspond with several theologians who had taken an

interest in the philosophical implications of his theories about the infinite. This correspondence was of special significance to Cantor because he was convinced that the transfinite numbers had come to him as a message from God [Dauben, 1977, 2005].

8 TRANSFINITE CARDINAL NUMBERS: THE ALEPHS (א)

Although not presented in the *Grundlagen*, one additional element of the technical development of transfinite set theory needs to be mentioned, as an important part of Cantor's on-going efforts to mount a convincing and satisfactory mathematical defense of his ideas: the transfinite *cardinal* numbers. The evolution of his thinking about the transfinite cardinals is curious. Although the alephs are probably the best-known legacy of his creation, they were the last part of his theory to be given either rigorous definition or a special symbol. Indeed, it is difficult in the clarity of hindsight to reconstruct the obscurity within which he must have been groping, and up to now his work has been discussed here largely as if he had already recognized that the power of an infinite set could be understood as a cardinal number.

In fact, beginning in the early 1880s, Cantor first introduced notation for his infinite (actually transfinite) sequence of derived sets P^{ν}, extending them well beyond the limitation he had earlier set himself to sets of the first species. However, at this time he only spoke of the indexes as 'infinite symbols' with no reference to them in any way as numbers.

By the time that Cantor wrote the *Grundlagen* in 1883, the transfinite ordinal numbers had finally achieved independent status as numbers, and were given the familiar omega notation, ω being the first transfinite ordinal number following the entire sequence of finite ordinal numbers, i.e. $1, 2, 3, \ldots, \omega$. However, there was no mention whatsoever in the *Grundlagen* of transfinite *cardinal* numbers, although Cantor clearly understood that it is the power of a set that establishes its equivalence (or lack thereof) with any other set, from which he would eventually develop his concept of transfinite cardinal number. But in the *Grundlagen*, he carefully avoided any suggestion that the power of an infinite set could be interpreted as a number.

Soon, however, Cantor began to equate the two concepts, especially in a lecture delivered in September 1883 to mathematicians at a meeting in Freiburg. Even so, no symbol was yet provided for distinguishing one transfinite cardinal number from another. Since he had already adopted the symbol 'ω' to designate the least transfinite ordinal number, when he finally introduced a symbol for the first transfinite cardinal number, it was borrowed from the symbols already in service for the transfinite ordinals. By 1886, in correspondence, Cantor had begun to represent the first transfinite cardinal as $\overset{*}{\omega}$, and the next largest he denoted $\overset{*}{\Omega}$; but this notation was not very flexible, and within months he realized the need for a more general notation capable of representing the entire ascending hierarchy of transfinite cardinals. Temporarily, he used fraktur $\mathfrak{o}$'s, derivatives from his omegas, to represent his sequence of cardinal numbers. For a time, he used an assortment of notations, including superscripted stars, bars, and his fraktur $\mathfrak{o}$'s interchangeably for transfinite cardinal numbers, without feeling any need to decide upon one or the other as preferable [Dauben, 1979, 179–183].

However, in 1893 the Italian mathematician Giulio Vivanti was preparing a general account of set theory, and Cantor realized it was time to adopt a standard notation. Only

then did he choose the Hebrew alephs (ℵ) for the transfinite cardinal numbers. He did so because he thought the familiar Greek and Roman alphabets were too common and already widely employed in mathematics for other purposes. His new numbers deserved something distinct and unique—and letters of the Hebrew alphabet were readily available among the type fonts of German printers. The choice of the alephs was particularly clever, as Cantor was pleased to admit, because the Hebrew aleph was also a symbol for the number one. Since the transfinite cardinal numbers were themselves infinite unities, the aleph could be taken to represent a new beginning for mathematics. Cantor designated the cardinal number of the first number class $\aleph_1$ in 1893, but in 1895 changed his mind; from then on, he used $\aleph_0$ to represent the first and least transfinite cardinal number, the number he had previously designated by $\overset{*}{\omega}$. From aleph-null, he went on to designate the cardinal number of the second number class as $\aleph_1$, after which there followed an unending sequence of transfinite cardinal numbers.

9 CANTOR AND THE *DEUTSCHE MATHEMATIKER-VEREINIGUNG*

During the 1880s Cantor had already begun to lay the strategic foundations for creating an independent union of mathematicians in Germany. The specific goal of such a union, as he often made clear in his correspondence, was to provide an open forum, especially for young mathematicians. The union (as Cantor envisaged it) would guarantee that anyone could expect free and open discussion of mathematical results without prejudicial censure from members of the older establishment, whose conservatism might easily ruin the career of an aspiring mathematician. This was primarily needed in cases where the ideas in question were at all new, revolutionary or controversial, as Cantor had learned from his experience with Kronecker.

Cantor labored intensively for the creation of the *Deutsche Mathematiker-Vereinigung* [Dauben, 1979, 160–163]. Eventually, agreement was reached and the Union of German Mathematicians held its first meeting in conjunction with the annual meeting of the *Gesellschaft Deutscher Naturforscher und Ärzte*, which met at Halle in 1891. Cantor was elected the Union's first president, and at its inaugural meeting he presented his now famous proof that the real numbers are nondenumerable using his new method of diagonalization [Cantor, 1891].

The Union was not the end of Cantor's vision. He also recognized the need to promote international forums as well, and thus began lobbying for international congresses shortly after formation of the Union. These were eventually organized through the cooperative efforts of many mathematicians, and not directly, it should be added, as a result of Cantor's exclusive efforts. The first of these was held at Zürich in 1897, the second in Paris in 1900 [Dauben, 1979, 163–165].

10 TRANSFINITE MATHEMATICS AND CANTOR'S MANIC DEPRESSION

Cantor made his last major contributions to set theory in his two-part paper *Beiträge* [Cantor, 1895–1897]. He had used his famous method of diagonalization in [Cantor, 1891] to show that the cardinal number of any set P is always less than the cardinal number of its

power-set, the set of all subsets of P. Now he presented a corollary to this result, namely that the cardinal number of the continuum is equal to $2^{\aleph_0}$, and he hoped this result would soon lead to a solution of the Continuum Hypothesis—which he could now express as $2^{\aleph_0} = \aleph_1$.

The arguments that Cantor used in his proof of 1891 about the cardinal number of the power-set of all subsets of any given set, however, led to far different conclusions. Rather than leading to a resolution of the Continuum Hypothesis, they led directly to discovery of the paradoxes of set theory, for the fact that there could be no 'largest' transfinite cardinal number immediately raised the question of the cardinality of the set of "all" transfinite cardinal numbers. Cantor resolved the problem by excluding this possibility entirely; the aggregate of "all' transfinite numbers was what he called an 'inconsistent' aggregate, and therefore was simply not to be considered as a 'set'. Bertrand Russell, in contemplating this problem, drew far more problematic conclusions, for what he discovered was that a paradox can be derived in set theory itself by considering those sets that do not include themselves as members (§61.1).

To understand Cantor's tenacious promotion and defense of set theory, especially in his later years after publication of the *Beiträge*, it is important to appreciate the connection between Cantor's faith in God, his mental illness, and his mathematics. Certain documents suggest that in addition to enforcing periodic intervals of contemplation and withdrawal from daily affairs, Cantor's periods of depression were not wholly unproductive. In fact, he was often able to pursue his mathematical ideas in the solitude of the hospital or quietly at home. This may all have supported his belief that the transfinite numbers had been communicated to him from God. In fact, as he noted in the third motto he chose to head the first part of his *Beiträge*, in 1895: 'The time will come when these things which are now hidden from you will be brought into the light'.

This is a familiar passage from the Bible, First Corinthians, and reflects Cantor's belief that he was an intermediary serving as the means of revelation. It may also be taken to reflect his faith that despite any prevailing resistance to his work, it would one day enjoy recognition and praise from mathematicians everywhere. Similarly, he considered the depressive phases of his bouts with manic depression to be periods during which he could devote himself to deep reflection, uninterrupted meditation, and even mathematics. Following a long period of hospitalization in 1908, he once wrote to the British mathematician Grace Chisholm Young (1868–1944), who then lived in Göttingen ([Meschkowski, 1971, 30–34]; translated in [Dauben, 1979, 290]):

> A peculiar fate, which thank goodness has in no way broken me, but in fact has made me stronger inwardly [...] has kept me far from home—I can say also far from the world [...] In my lengthy isolation neither mathematics nor in particular the theory of transfinite numbers has slept or lain fallow in me.

Elsewhere, Cantor actually described his conviction about the truth of his theory explicitly in quasi-religious terms. For example, in a letter of 1888 [Dauben, 1979, 298] he wrote:

> My theory stands as firm as a rock; every arrow directed against it will return quickly to its archer. How do I know this? Because I have studied it from all

sides for many years; because I have examined all objections which have ever been made against the infinite numbers; and above all, because I have followed its roots, so to speak, to the first infallible cause of all created things.

11 CONSEQUENCES OF THE *GRUNDLAGEN* FOR LATER MATHEMATICS

Some of the immediate and long-term results of Cantorian set theory are reflected in a number of works included as 'Landmarks' in this volume, notably those by Hilbert, Henri Lebesgue and René Baire, and Russell and A.N. Whitehead. Additionally, among the most immediate results of Cantor's new theory of transfinite ordinal numbers were applications in real and complex analysis, for example, Mittag-Leffler's theorem of 1884 that extended results obtained by Weierstrass on analytic representations of complex functions to meromorphic functions. For references for this section, see the superb bibliography in [Fraenkel, 1953].

In France, Emile Borel's doctoral thesis of 1894 used set theory in applications to problems of analytic continuation in the theory of complex functions. Among classic results that Borel reached using transfinite ordinals was the Heine–Borel covering theorem. At the International Congress of Mathematicians in 1897 in Zurich, Adolf Hurwitz treated analytical functions in conjunction with transfinite ordinals and used them to classify analytical functions. Jacques Hadamard gave a general lecture on future applications of set theory, suggesting that sets of functions were especially intriguing, and predicting that in partial differential equations and mathematical physics, set theory might be especially useful. Also significant in exploiting set theory in analysis was Maurice Fréchet's thesis of 1906, 'Sur quelques points du calcul fonctionnel'.

Mathematicians like Baire, Fréchet, Hilbert, and Marcel Riesz all drew on set-theoretic notions to develop increasingly abstract concepts of space, as outlined for example in results on 'Stetigkeit und abstrakte Mengenlehre' that Riesz presented at the International Congress of Mathematicians in Rome in 1908. Finally here, in his book *Grundzüge der Mengenlehre* (1914) Felix Hausdorff included significant new results on order-types, and topological and metric spaces.

In England, the most important early contributions to develop Cantorian set theory were made by Russell and Whitehead, P.E.B. Jourdain, Grace Chisholm and William Henry Young (largely from abroad) and G.H. Hardy. Cantor's theory of transfinite numbers has been further extended into the realm of inaccessible cardinals and a host of other highly refined theories of transfinite numbers. Finally, the legacy of Cantor's transfinite set theory that the *Grundlagen* launched in 1883 has been especially important for the development of mathematical logic and work on foundations related to the paradoxes of the infinite, logical paradoxes, and work associated with Ernst Zermelo, Bertrand Russell, Abraham Fraenkel, Kurt Gödel (§71) and Paul Cohen, among many others.

BIBLIOGRAPHY

Charraud, N. 1994. *Infini et inconscient. Essai sur Georg Cantor*, Paris: Anthropos.

Cantor, G. *Papers. Gesammelte Abhandlungen mathematischen und philosophischen Inhalts* (ed. E. Zermelo), Berlin: Springer, 1932. [Repr. Hildesheim: Olms, 1966; Berlin: Springer, 1990.]

Cantor, G. 1872. 'Über die Ausdehnung eines Satzes aus der Theorie der trigonometrischen Reihen', *Mathematische Annalen*, *5*, 123–132. [Repr. in [*Papers*, 92–102]. French trans. in *Acta mathematica*, *2* (1883), 336–348.]

Cantor, G. 1874. 'Über eine Eigenschaft des Inbegriffes aller reellen algebraischen Zahlen', *Journal für die reine und angewandte Mathematik*, *77*, 258–262. [Repr. in [Cantor, *Papers*, 115–118]. French trans. in *Acta mathematica*, *2* (1883), 205–310.]

Cantor, G. 1879. 'Über unendliche, lineare Punktmannigfaltigkeiten', Part I, *Mathematische Annalen*, *15*, 1–7. [Repr. in [Cantor, *Papers*, 139–145]. French trans. in *Acta mathematica*, *2* (1883), 349–356.]

Cantor, G. 1883. 'Fondements d'une théorie générale des ensembles', *Acta mathematica*, *2*, 381–408.

Cantor, G. 1891. 'Über eine elementare Frage der Mannigfaltigkeitslehre', *Jahresbericht der Deutschen Mathematiker-Vereinigung*, *1*, 75–78. [Repr. in [Cantor, *Papers*, 278–280].]

Cantor, G. 1895–1897. 'Beiträge zur Begründung der transfiniten Mengenlehre', *Mathematische Annalen*, *46*, 481–512; *49*, 207–246. [Repr. in [Cantor, *Papers*, 282–356]. English trans. in *Contributions to the founding of the theory of transfinite numbers* (ed. and trans. P.E.B. Jourdain), La Salle, Ill.: Open Court, 1915 (repr. New York: Dover, 1955).]

Cantor, G. and Dedekind, R. 1937. *Briefwechsel Cantor–Dedekind* (eds. E. Noether and J. Cavaillès), Paris: Hermann.

Dauben, J.W. 1971. 'The trigonometric background to Georg Cantor's theory of sets', *Archive for history of exact sciences*, *7*, 181–216.

Dauben, J.W. 1977. 'Georg Cantor and Pope Leo XIII: mathematics, theology, and the infinite', *Journal of the history of ideas*, *38*, 85–108.

Dauben, J.W. 1979. *Georg Cantor: his mathematics and philosophy of the infinite*, Cambridge, MA: Harvard University Press. [Repr. Princeton: Princeton University Press, 1990.]

Dauben, J.W. 2005. 'Georg Cantor and the battle for transfinite set theory', in *Kenneth O. May Lectures of the Canadian Society for History and Philosophy of Mathematics*, Berlin: Springer, to appear.

Dedekind, R. 1872. *Stetigkeit und irrationale Zahlen*, 1st ed., Braunschweig: Vieweg. [English trans. in *Essays on the theory of numbers* (trans. W.W. Beman), Chicago: Open Court, 1901 (repr. New York: Dover, 1963), 1–27. See §43.]

Edwards, H.M. 1989. 'Kronecker's views on the foundations of mathematics', in David E. Rowe and John McCleary (eds.), *The history of modern mathematics*, Boston, MA: Academic Press, vol. 1, 67–77.

Ewald, W.B. 1996. (Ed.), *From Kant to Hilbert: A source book in the foundations of mathematics*, 2 vols., New York: Oxford University Press.

Fraenkel, A. 1930. 'Georg Cantor', *Jahresbericht der Deutschen Mathematiker-Vereinigung*, *39*, 189–266.

Fraenkel, A. 1953. *Abstract set theory*, Amsterdam: North-Holland.

Grattan-Guinness, I. 1971. 'Towards a biography of Georg Cantor', *Annals of science*, *27*, 345–391.

Heine, E.H. 1870. 'Über trigonometrische Reihen', *Journal für die reine und angewandte Mathematik*, *71*, 353–365.

Meschkowski, H. 1967. *Probleme des Unendlichen. Werk und Leben Georg Cantors*, Braunschweig: Vieweg.

Meschkowski, H. 1971. 'Zwei unveröffentlichte Briefe Georg Cantors', *Der Mathematikunterricht*, *4*, 30–34.

Moore, G.H. 1982. *Zermelo's axiom of choice*, New York: Springer-Verlag.

Purkert, W. and Ilgauds, H.J. 1985. *Georg Cantor*, Leipzig: Teubner.

Purkert, W. and Ilgauds, H.J. 1987. *Georg Cantor 1845–1918*, Basel: Birkhäuser, 1987.

Schoenflies, A. 1927. 'Die Krisis in Cantor's mathematischem Schaffen', *Acta mathematica*, *5*, 1–23.

Thuillier, P. 1977. 'Dieu, Cantor et l'infini', *La recherche*, *84*, 1110–1116.

CHAPTER 47

RICHARD DEDEKIND (1888) AND GIUSEPPE PEANO (1889), BOOKLETS ON THE FOUNDATIONS OF ARITHMETIC

J. Ferreirós

These works offered new levels of foundation and formalisation for arithmetic, stressing particularly the understanding of the method of mathematical induction. They established the Peano–Dedekind axioms for the natural numbers.

Dedekind

First publication. Was sind und was sollen die Zahlen?, Braunschweig: Friedrich Vieweg & Sohn, 1888. 54 pages. [Actually issued December 1887.]

Manuscripts. Main manuscript destroyed, but first draft (written 1872–1878) preserved at the *Niedersächsische Staats- und Universitätsbibliothek Göttingen, Handschriften-abteilung*, Cod. Ms. Dedekind, III, 1: transcribed in [Dugac, 1976, 293–309]. Also fragments of a second draft of 1887 [*ibidem*, III, 1, III].

Later editions. 2nd ed. 1893 (new preface), 3rd ed. 1911 (new preface). Both Vieweg. [Various photoreprints until the so-called '10th ed.' 1969, prefaced by G. Asser (itself repr.). Also in *Gesammelte mathematische Werke*, vol. 3, Braunschweig: Vieweg, 1932, 335–391.]

English translations. 1a) 'The nature and meaning of numbers' (trans. W.W. Beman), in R. Dedekind, *Essays on the theory of numbers*, Chicago, Open Court, 1901, 29–115. [Several reprs. inc. New York: Dover, 1963.] 1b) Revised by W. Ewald in his (ed.), *From Kant to Hilbert*, New York: Oxford University Press, 1996, 790–833. 2) *What are numbers and what should they be?* (ed. and trans. H. Pogorzelski, W. Ryan and W. Snyder), Orono, Maine: Research Institute for Mathematics, 1995.

Russian translation by N.N. Parfentev in *Izvestija fiziko-matematicheskogo obtshchestva Kazan*, 15 (1905), 25–103.

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

Italian translations. 1) In *Essenza e significato dei numeri* (trans. with historico-critical notes by O. Zariski), Rome: Stock, 1926, 1–18. 2) In R. Dedekind, *Scritti sui fondamenti della matematica* (trans. F. Gana), Naples: Bibliopolis, 1982.

Japanese translation. In *On number—continuity and the nature of number* [in Japanese] (trans. Kono Isaburo), Tokyo: Iwanami Shoten, 1961.

French translation. Les nombres: *que sont-ils et à quoi servent-ils?* (trans. I. Milner and H. Sinaceur, introd. by M.-A. Sinaceur), Paris: La Bibliothèque d'Ornicar, 1978.

Spanish translation. ¿Qué son y para qué sirven los números? (trans. and introd. J. Ferreirós), Madrid: Alianza Editorial/UAM, 1998.

Peano

First publication. Arithmetices principia, Turin: Bocca, 1889. 16 + 20 pages.

Later edition. G. Peano, *Opere scelte*, vol. 2 (ed. and introd. U. Cassina), Rome: Cremonese, 1958, 20–55.

English translations. 1) Partial by J. van Heijenoort as 'The principles of arithmetic, presented by a new method', in his (ed.), *From Frege to Gödel*, Cambridge, MA; Harvard University Press, 1967, 83–97. 2) Full in *Selected works of Giuseppe Peano* (trans. and ed. H.C. Kennedy), London: George Allen & Unwin, 1973, 101–134.

Bilingual Latin–Spanish edition. *Arithmetices principia*, Oviedo: Pentalfa, 1979.

Related articles: Grassmann (§32), Dirichlet (§37), Dedekind on irrational numbers (§43), Boole (§36), Riemann on trigonometric series (§38), Cantor (§46), Whitehead and Russell (§61).

1 DEDEKIND: BIOGRAPHY AND BACKGROUND

Richard Dedekind was born in 1831 in Braunschweig (Brunswick) and educated at the Collegium Carolinum there (an institution offering courses of a level intermediate between secondary school and university) before entering the University of Göttingen. He was following in the footsteps of the great C.F. Gauss (1777–1855), of whom he would become the last doctoral student in 1852. Similarities can also be found at the personal level: both men were sober and upright, somewhat shy; both followed the motto 'pauca sed matura', although no doubt Gauss was more prolific.

Richard was the fourth child of a well-to-do family, son of a lawyer and professor at the Carolinum. During the student years at Göttingen he stood out in social life as a piano and cello player, but he also enjoyed Wilhelm Weber's lectures in physics. It is unlikely that Gauss that was a powerful influence by personal contact, but he certainly was through his classic writings. Indeed, the direction of Dedekind's lifework had been set by the great German tradition in algebraic number theory, inaugurated by Gauss (§22) and significantly advanced by Ernst Kummer.

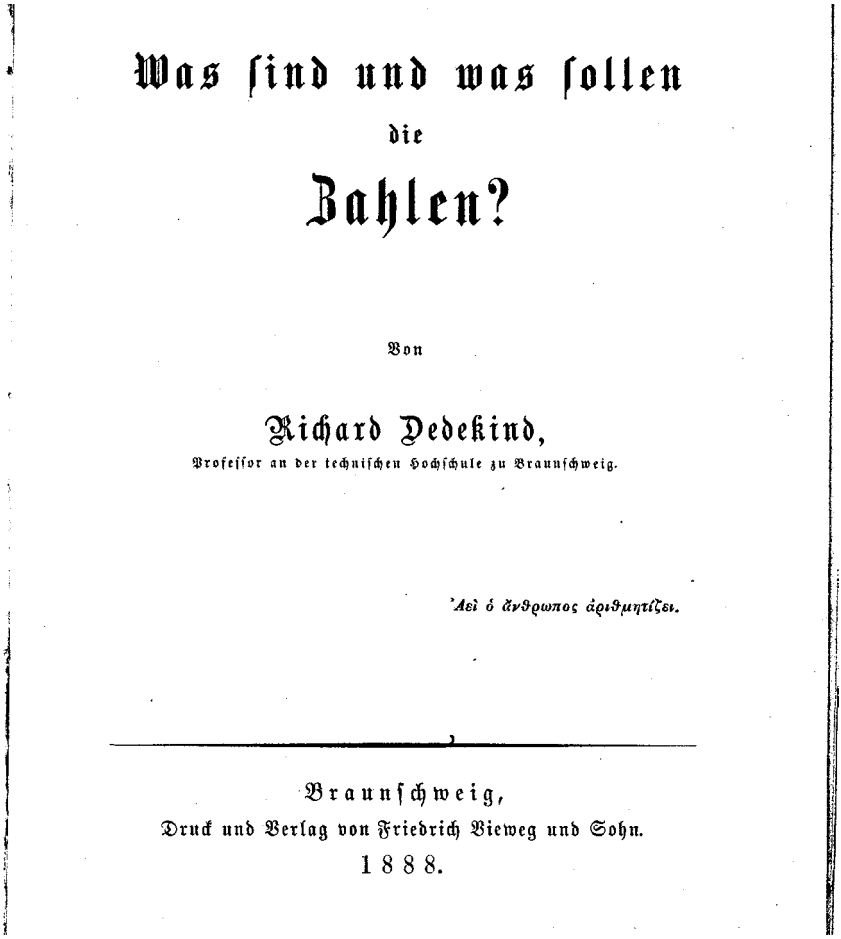
According to Dedekind himself, it was only after his *Habilitation* as *Privatdozent* in 1854 that he had the experience of entering a high-level research school. His new masters

were J.P.G. Lejeune-Dirichlet, who followed a call to Göttingen upon the death of Gauss in 1855 (at whose funeral Dedekind was a pall-bearer), and his fellow *Privatdozent* Bernhard Riemann (only 5 years older). Dedekind attended all of their lectures, and some years later he would see through the posthumous publication of some or all of their work. The influence of Dirichlet was definitive at the level of rigour, mathematical proof, and the effort to fill gaps in Richard's mathematical education. That of Riemann would be particularly important at the level of abstract methodology and the turn toward a set-theoretical approach ([Ferreirós, 1999]: compare §46).

In 1858 Dedekind accepted a call from the Zurich Polytechnical School, and it was during his first year in Zurich that he formulated his well-known definition of the real numbers. By then his main work was in algebraic number theory, but he also elaborated ideas on the foundations of the number concept. In 1862 he moved back to Braunschweig, accepting a chair at the *Collegium Carolinum*, which soon turned into a Technical School, partly under Dedekind's guidance in the role of rector. He never accepted calls from any university, and thus he did not have noteworthy students; nor did he found a school. In 1881 Georg Cantor (1845–1918), with whom he had a remarkable correspondence, tried to win him for Halle, but even this attempt failed. Dedekind never married, living with his mother and his sister Julie, a successful writer. He became emeritus professor in 1896, and he died in the middle of the Great War, in February 1916.

Some time before 1872, Dedekind's views on the number concept ripened to the point of seeing the technical possibility of reducing the whole number system (and thus, in his view, all of pure mathematics) to the basic concepts of set and mapping. This confirmed him in the logicist beliefs that he is likely to have entertained since the successful reduction, in 1858, of the real numbers to rationals plus sets. This significant train of thought followed along the lines of previous ideas of Gauss and Riemann (with noteworthy parallels to M. Ohm, Hermann Grassmann (§32) and Karl Weierstrass), but it went considerably further. Between 1872 and 1878 he wrote down a series of draft notes which contain all the most significant and enduring ideas that would be published in *Was sind und was sollen die Zahlen?* (hereafter, '*WSZ*'). A remarkable reconstruction of the path leading to this work is contained in Dedekind's letter of 1890 to a school-teacher, Hans Keferstein (published in translation in [van Heijenoort, 1967, 98–103]. The booklet appeared in 1888 (Figure 1), when he was in his fifties; for a portrait from around that time, see Figure 2.

In the words of Felix Klein, who obviously had a different character, Dedekind was 'a contemplative nature': he was a system builder. A key feature of his work is that he kept constantly intertwining the concepts and methods of his more advanced research, with those in his foundational writings. Symptomatic is the fact that he first announced the publication of *WSZ* in the second version of his ideal theory in the *Vorlesungen über Zahlentheorie* ([Dirichlet and Dedekind, 1879]: see §37), and that the third version [Dirichlet and Dedekind, 1894] contains a good number of footnotes referring back to *WSZ*. One may say that *WSZ* is linked, in one way or another, to all of Dedekind's research topics, including even some ideas that he left unpublished on the foundations of topology and of projective geometry.



Was ſind und was ſollen
die
Zahlen?

Von

Richard Dedekind,
Profeſſor an der techniſchen Hochſchule zu Braunſchweig.

Ἀεὶ ὁ ἄνθρωπος ἀριθμητίζει.

Braunſchweig,
Druck und Verlag von Friedrich Vieweg und Sohn.
1888.

Figure 1. Title page of Dedekind's booklet in its first edition. The Greek motto reads 'aei ho anthropos arithmetidsei' ('man is constantly involved in arithmetic'). It constitutes Dedekind's reply to Gauss and Plato ('God is always involved in arithmetic', or 'geometry').

2 PEANO: BIOGRAPHY AND BACKGROUND

Giuseppe Peano (Figure 3) was 27 years younger than Dedekind. Born in August 1858 in Spinetta (province of Cuneo, in the Piedmont), Peano was the son of a peasant. He studied at the University of Turin from 1876, taking his doctoral degree in 1880, and remained there until his death in 1932, working first as an assistant, then from 1890 as extraordinary professor, finally from 1895 as full professor. During the 1880s he worked in analysis, probably containing his most important results; particularly noteworthy are the continuous space-filling Peano curve (1890), and the notion of content (a precedent of measure theory)

Figure 2. Portrait of Dedekind in his fifties.

developed in the 1880s independently of Camille Jordan. The textbook that he published in 1884, based on the lectures of his teacher Angelo Genocchi (*Calcolo differentiale e principii di calcolo integrale*), already contained a great number of new results of his own.

The years 1889 to 1908 saw Peano's intensive dedication to symbolic logic, the axiomatization of mathematics, and the production of the encyclopedic *Formulaire de mathématiques* (5 editions, 1895–1908). This was an ambitious assembly of mathematical results,

Figure 3. Portrait of Peano in the 1880s.

compactly presented with the symbolic means of logic, though completely without proofs. In 1891 he founded with some colleagues the journal *Rivista di matematica*, gathering around him an important group of followers. Peano was an accessible man, and the way in which he mingled with students was regarded as 'scandalous' in Turin. He was a socialist in politics, and a tolerant universalist in all matters of life and culture.

Like other logicians of the time, Peano was strongly interested in the Leibnizian dreams of a *characteristica universalis* and a *calculus ratiocinator*. His development of a powerful, clear, and flexible logical language was a step in this direction. In the late 1890s, he became increasingly interested in elaborating a universal spoken language. The outcome was *Latino sine flexione*, based on Latin and incorporating vocabulary from the main European languages, with no grammatical inflexions. The last edition of the *Formulario* (1905–1908) appeared in this language.

Peano followed closely the work of German mathematicians such as Grassmann, Ernst Schröder and Dedekind; for example, the 1884 textbook presented a definition of the real numbers by Dedekind cuts, and Peano also studied the *Vorlesungen über Zahlentheorie*. In 1888 he published *Calcolo geometrico secondo l'Ausdehnungslehre di H. Grassmann*, containing a study of the operations of deductive logic, which was to be refined and extended in later years. In 1889 he presented—notably in Latin—a first version of the famous axioms for **N**, which he refined in volume 2 of the *Formulaire* [1898a].

Peano's work on the natural numbers was at the crossroads of his diverse mathematical contributions, linking naturally his previous research in analysis with the coming focus on logical foundations, and being a necessary prerequisite for the *Formulaire* project. It aimed at filling the most significant gap in the foundations of mathematics at a time when the 'arithmetization' of analysis had essentially been completed. It is no mere coincidence that other mathematicians (Gottlob Frege, C.S. Peirce and Dedekind) published convergent work in the same decade. As far as we know, all of these men worked independently. In comparison with the rest, Peano's attempt is better rounded than Peirce's, but shallower than those of Frege and Dedekind. Partly because of this last trait, it has been more popular.

In the preface to *Arithmetices principia*, Peano stated that, for proofs in arithmetic, he was relying on the handbook [Grassmann, 1861], and that the 'acute' work of Dedekind 'was also most useful' to him (p. vi). However, according to Peano himself, his booklet was already in print before he first saw *WSZ*, and Dedekind's work merely provided 'moral proof of the independence' of his axioms [Peano, 1898b, 243]. This seems plausible also for internal reasons. Had he studied thoroughly Dedekind's work before writing his own, Peano would probably have published a different, deeper study of the subject. As things happened, under time pressure, he remained content with verifying that Dedekind's analysis agreed with his own.

Actually, *Arithmetices principia* can be regarded as a simplification, refinement, and translation into logical language (the '*nova methodo*' in its title) of Grassmann's *Lehrbuch der Arithmetik* [1861]. Grassmann strived to elaborate a stern deductive structure and he stressed proofs by mathematical induction and recursive definitions. But curiously, unlike Peano, he did not postulate an axiom of induction. Thus Peano presented more clearly the basic assumptions. He also simplified the system by starting with the natural numbers and not the integers, as did Grassmann. Grassmann's work was unknown to Dedekind as late as 1876 (see Lipschitz [1986, 74]); in any case, Dedekind's early work on **N** around 1860

and his first (1863) edition of the *Vorlesungen* (§37) already emphasized mathematical induction and recursive definitions.

3 DEDEKIND'S THEORY

As Hilbert once said, Dedekind elaborated an 'extremely sagacious' construction of the natural numbers—he had the 'dazzling and captivating' idea of grounding the finite numbers on the actual infinite [Hilbert, 1905, 1922]. This parallels what happens in his famous definition of infinity (*WKZ*, art. 64), which contrasting with all previous analyses defines the finite as the non-infinite, and establishes that a set S is (Dedekind-)*infinite* if and only if there exists a bijective mapping of S onto a proper part of itself.

The contents of Dedekind's book are summarised in Table 1. He realized that a general theory of sets and mappings is a sufficient foundation for the natural numbers, and indeed the usual set-theoretic definitions of N exemplify his ideas. Thus the first two sections of *WSZ* are devoted to '*Systeme*' (his technical term for sets, equivalent to Cantor's '*Mengen*') and to mappings ('*Abbildungen*'). Although he presented basic results very clearly and succinctly, Dedekind lacked terminology and notation for the membership relation. Indeed, he exploited his notation for inclusion to the point of making it equivocal, using $a \subset S$ as shorthand for $\{a\} \subset S$ and thus for $a \in S$. (Instead of '$\subset$' he actually used a symbol that looks like '3'.) This would be severely criticized by Frege, but, as a matter of fact, letters and manuscripts written in 1888 and 1889 show that Dedekind never was unclear as to the distinction between inclusion and membership, and that he fully realized the 'dangers' (his term) implicit in his equivocal usage of notation.

WSZ was the first work to thematise the concept of mapping and to discuss its basic theory. This makes the more noteworthy that Dedekind was able to handle it in such a masterful way. But in fact that was an instance of 'pauca sed matura': the concept had been in use (under a different name) in all of Dedekind's algebraic work, and it can be traced

Table 1. Contents by chapters of Dedekind's book. 54 pages.

Chs.	**Arts.**	**(Short) titles or descriptions**
Preface		
1–2	1–25	Systems of elements; mapping of a system.
3–4	26–63	Similarity of a mapping; similar systems.
5–6	64–80	Finite and infinite; sequence of natural numbers.
7	81–118	Larger and smaller numbers.
8	119–123	Finite and infinite parts of the number sequence.
9	124–131	Definition of a mapping of the number sequence by recursion ['Induktion'].
10	132–134	The class of simply infinite systems.
11–13	135–158	Addition, multiplication, exponentiation of numbers.
14	159–172	(Cardinal) number of the elements of a system.

back to the late 1850s [Ferreirós, 1999, 88–99]. Also very modern and mature is, e.g., the treatment of equipollence as an equivalence relation, but again this had a long prehistory in his algebraic and number-theoretical work. The only shortcoming is that Dedekind did not differentiate clearly between injectivity and bijectivity: he defined the first concept but in fact he frequently employed the second. (Maybe he found it trivial and innocuous to restrict the final set to the image.) A bijective mapping is called 'similar' or 'distinct'—natural terms to use in conjunction with *Abbildung*, a word that means 'representation'. Such a term suggests a philosophical turn, a theme discussed by Dedekind himself in the first preface to *WSZ*.

The most original and profound theoretical development in *WSZ* consists in the so-called theory of *chains*, elaborated in a section on internal mappings (ch. 4). Having seen that the structure and operations of N can be characterized purely in terms of the successor mapping σ, Dedekind noticed three essential traits of this structure: the mapping $\sigma : N \to N$ is internal, it is also injective or 'similar', and there is a 'base element' (denoted 1) such that $1 \notin \sigma(N)$. These three conditions were enough to prove that N is Dedekind-infinite. Nevertheless, as he explained in the letter to Keferstein mentioned above, Dedekind soon realized that the three conditions would also be met by deviant (non-standard) sets N^* containing unwelcome extra elements, possibly in such a way that proofs by mathematical induction would not succeed in establishing the result for all members of N^*. (The 1870s draft shows traces of this noteworthy argument [Dugac, 1976, 295], which is a forerunner of model theory.) A fourth condition was needed to fully characterize N, and this would be formulated by means of the concept of a chain.

Thus, the concept of *chain of a subset* was obtained by analyzing and generalizing the conditions that an internal mapping must satisfy in order to make proofs by induction possible. Given $\varphi : S \to S$ and $A \subset S$, the chain of subset A is the intersection of all subsets K of S such that $A \subset K$ and $\varphi(K) \subset K$. This is the closure of A under φ in S, the smallest subset of S containing A and closed under φ, denoted by $\varphi_0(A)$. Now, it suffices to add a fourth condition to the three ones given in the last paragraph: N is the chain of $\{1\}$, i.e., $N = \sigma_0(\{1\})$. The four conditions characterize the structure of what Dedekind calls a *simply infinite set*, and they turn out to be equivalent to Peano's axioms. In particular, the key chain condition is Dedekind's sophisticated version of the axiom of mathematical induction (axiom 9 below; the first three conditions are equivalent to the Peano axioms 1, 6, 7, 8).

Dedekind was able to show that all simply infinite sets are isomorphic, that is, that his characterization of N was categorical or monomorphic (ch. 10: the result can only be recovered within a second-order logical framework). On the basis of the concept of chain, he studied the relations $<$ and $\leqslant$, and derived basic results on finite and infinite subsets of N. Most notably, he was the first to offer a general theory and justification of recursive definitions (ch. 9: 'mappings of the number-series defined by induction') and to study primitive recursive functions. He proved the following theorem (art. 126): given any mapping θ of set Ω in itself (injective or not), and given an element $\omega \in \Omega$, there is one and only one mapping $\psi : N \to \Omega$ such that $\psi(1) = \omega$ and $\psi(n') = \theta\psi(n)$, where n' is the successor of n.

Then Dedekind proceeded to introduce by recursive definition the operations of addition, multiplication, and powering, establishing their properties in a perfectly rigorous

way. Furthermore, he employed the theorem of art. 126 as a basis to justify the introduction of the finite cardinal numbers (ch. 14). To do so, he offered a proof that a set is (Dedekind-)finite if and only if there exists an initial segment $Z_n = \{x\colon\ x \in N \text{ and } x < n\}$ equipollent to the given set. In his approach (art. 159), this required to show that, if S is an infinite set, then each initial segment Z_n can be mapped onto S, and conversely. Dedekind remarked that proving the converse, evident as it might seem, was 'complicated', and in fact his proof relied implicitly upon the axiom of choice.

One of the distinguishing characteristics of *WSZ* is that Dedekind proceeded to investigate the required concepts in great generality. While all he needed for his limited topic was chains of an element under a bijective mapping (as the sets $N \setminus Z_n$ always are), his theory of chains studied chains of arbitrary sets under arbitrary mappings. In this way, he obtained an extremely useful tool for the development of set theory, but he failed to spell this out for his readers. And here we find a little mystery within *WSZ*: the exposition of chain theory (ch. 4) ends with a proposition that establishes (in generalized form) the crucial lemma for proving the Cantor–Bernstein theorem, a lemma that Cantor himself had formulated.

The Cantor–Bernstein theorem is a very basic result in the theory of cardinalities: if set A is equipollent to a part of B, and B is equipollent to a part of A, then A and B are themselves equipollent. Dedekind was aware of the importance that this result had for his correspondent Cantor and for general set theory. Cantor searched for a proof in vain since 1882, but his student Felix Bernstein succeeded in 1897. Dedekind's proof is simpler and more elegant, and he obtained it during 1887 while preparing the final draft for *WSZ* (see *Werke*, vol. 3, 447–448). It seems clear that it was not inadvertently that he presented the crucial lemma in art. 63, leaving the proof for his readers. But Cantor failed to appreciate the implications and the real scope of Dedekind's results in chain theory [Ferreirós, 1993; 1999, 239–241].

4 PEANO'S THEORY

The contents of Peano's booklet *Arithmetices principia* are summarised in Table 2. His main aim was to elaborate a perspicuous logical notation, adequate to the goal of symbolically rewriting all known mathematics. In his view, this would enable him to avoid 'the ambiguity of language' and would thus provide the key to a satisfactory solution of

Table 2. Contents by chapters of Peano's booklet.

Page(s)	**Chs.**	**(Short) titles or descriptions**
iii–xvi		Preface: notations of logic.
1	1–2	Numbers and addition; subtraction.
6	3–6	Maxima and minima; addition; multiplication; powers; division.
12	7	Various theorems; rational numbers.
14	8–9	Rational and irrational numbers.
17	10	Systems of real numbers. [End 20.]

the question of the foundations of mathematics. The first version of his axiom system intermingled the principles of the logic of identity (restricted to number objects—axioms 2 to 5) with the famous five axioms (p. 34). To facilitate reading, I have slightly modernized Peano's logical notation, replacing some dots by parentheses and employing the symbol '$\wedge$' for conjunction ('and'):

1. $1 \in N$.
6. $a \in N. \supset .a + 1 \in N$.
7. $a, b \in N. \supset .(a = b. = .a + 1 = b + 1)$.
8. $a \in N. \supset .a + 1 - = 1$.
9. $[k \in K \wedge 1 \in k \wedge x \in N \wedge (x \in k. \supset .x + 1 \in k)]. \supset .N \supset k$.

The second occurrence of '$=$' in 7. denotes 'if and only if' and would now be rendered by '$\leftrightarrow$'; in 8., '$- =$' means '$\neq$'; the last instance of '$\supset$' in 9. denotes set inclusion.

'$1 \in N$' was read, alternatively, '1 is a number' and '1 belongs to (the class) N'. As one can readily notice, the logical symbolism was systematically given a dual reading, on the one hand as elementary logical operations, on the other as basic set-theoretical operations. (The tradition of doing so began with Boole 50 years earlier: see §36.) Moreover, the principle of comprehension was presupposed in Peano's notation. As is well known, this principle was responsible for the contradictions, paradoxes or antinomies discovered around 1900 (§61.1). The contradictory principle of comprehension was also implicit in Dedekind's work (see especially *WSZ*, art. 60). Related to all of this, Peano's axioms are meant to be formalized in second-order logic, not the first-order logic which we usually employ since the time of Hilbert's program.

In later presentations Peano [1898a] reduced his system to five axioms, treating the principles of identity as logic, and he chose to take 0 as the base element. Meanwhile, Peano [1891] offered five 'primitive propositions' that are 'due to Dedekind' (p. 86; see p. 84), though at first sight they seem just a modification of Peano's original system. (He admitted that he had altered the chain principle β, bringing it closer to his own induction principle; he seems never to have grasped the great generality of Dedekind's theory of chains.) On the basis of that declaration, it is frequently stated that Peano simply took his axioms from Dedekind, and acknowledged to have done so in [1891].

But this interpretation is contrary to Peano's own statements, and it overlooks the differences between the system in [Peano, 1891] and those in this booklet and [1898a]. Closer scrutiny reveals that in [Peano, 1891] he formulates the axioms as conditions on a set N, its subsets, and a successor function. Meanwhile, when the axioms are presented in the other two works they are formulated more elementarily as conditions on the *elements* of a set. This seems to be the way Peano himself understood the main difference between his analysis and that of Dedekind (see also [1891, 88] and [1898b, 243]).

It deserves to be mentioned that, like [Grassmann, 1861], Peano's *Arithmetices principia* not only deals with the natural numbers. Having presented all of the basic arithmetical operations, Peano goes on to discuss several topics in order to 'better show the power of this [new] method'. First he offers a selection of number-theoretical results (without proof, art. 7) and then he introduces the rational and the real numbers. The rationals are rendered

as ratios of two natural numbers (art. 8), the reals as formal 'limits' of sets of rationals, essentially following Dedekind's definition by means of cuts (art. 9): the treatment of the real numbers was the most unsatisfactory aspect of [Grassmann, 1861]. Finally, art. 10 discusses basic results in the topology of the real numbers, belonging to the theory of 'what Cantor calls *Punktmenge* (*ensemble de points*)', on the basis of the concepts of interior, exterior and limit point. Some of these results were new; many years earlier, in unpublished work unknown to Peano, Dedekind had also defined the concepts of interior, exterior and boundary [Ferreirós, 1999, 138–139].

As we see, at first sight Peano's treatise may look more modern than Dedekind's, because he employs the language of logic, and with notations that have been extremely influential. Nevertheless, Dedekind's account of the foundations of the natural number system was much more detailed. An example is the issue of recursive definitions. Peano uses them, like Grassmann and others, although they do not comply with his own explicit restrictions on definitions. Only in Dedekind we find a general account and justification of recursive definitions (art. 126; see above), established as a foundation for the introduction of the arithmetical operations. Likewise, in Dedekind we find detailed arguments to the effect that the chosen axioms are actually sufficient to characterize N. And, while Peano's axiom of induction is a direct rendition of the customary principle, Dedekind subsumed it under the very general and powerful theory of chains.

Those same reasons have the side effect that Dedekind's work becomes more difficult for a general reader than Peano's. There is, finally, one important difference between both works, which this time makes Peano's clearly closer to us: while he uses clearly and very explicitly the modern axiomatic terminology, Dedekind carefully avoided the very word 'axiom'. As to the reason for this, in my view that was because of the logicist project, the way that Dedekind (and Frege too) understood it. The defining conditions of N were not axioms, but perfectly legitimate logical stipulations, from which all relevant arithmetical results follow. In particular, all existential presuppositions would have been previously guaranteed on the basis of logic alone. The one existential assumption that Dedekind was unable to justify, in spite of his good efforts inspired by Bolzano, is the axiom of infinity [Ferreirós, 1999, 244–248, 252–253].

5 APPRAISAL AND IMPACT

While even today some authors regard Dedekind's approach as 'formal' and too abstract, others welcome it as a key instance of structural reasoning in mathematics. In spite of his critical attitude toward the set-theoretical approach, Hermann Weyl stated that Dedekind's booklet had marked an epoch in the development of mathematical thought [Weyl, 1918, 16; also pp. 35–36]. This seems to have been a Göttingen theme, probably fostered by Hilbert himself (compare §55). In his crucial paper on the axiomatization of set theory, Ernst Zermelo [1908, 200] referred to it as the 'theory created by Cantor and Dedekind'. While his early work in set theory had focused only on Cantor's work, from 1905, probably under the influence of Hilbert, Zermelo had made a close study of Dedekind's contributions.

The impact of *WSZ* has been a matter of some contention during recent years. When it was published, Dedekind was already very famous as a number theorist, and his booklet

received considerable attention. Peano referred to it while presenting his axiom system for N, remarking on the essential equivalence of their characterizations (p. vi). Subsequently he offered an elaboration of Dedekind's views in his own symbolism [Peano, 1891], and in the *Formulaire* he reproduced excerpts from Dedekind's work in the original German, reviewing its most important steps [1898a, 218–219]. Indeed, followers of Peano such as Rodolfo Bettazzi and Cesare Burali-Forti relied extensively on Dedekind's work during the 1890s.

Hilbert paid attention to *WSZ* immediately after its appearance, and in 1891 he would refer to it in his lectures, apparently endorsing Dedekind's views [Hallett and Majer, 2004]. Frege commented on the work in detail in his *Grundgesetze*, critically as usual; but he acknowledged that it was the most complete and noteworthy recent contribution to the foundations of mathematics [Frege, 1893, vii]; on Frege's critical stance, see [Tait, 1996].

Schröder included an enthusiastic detailed review of *WSZ* in volume 3 of his lectures on logic. He made it one of the 'most important objectives' of his work to incorporate Dedekind's essential ingredients (especially the theory of mappings and chains) into general logic, namely under the theory of relations [Schröder, 1895, 346–352]. Indeed, lectures 9 and 12 in this volume deal with Dedekind's most general ideas. Bertrand Russell read Dedekind's work long before his acquaintance with Peano and Frege [Garciadiego, 1992]; while his reaction to important aspects of it was negative, perhaps the topic of Dedekind's role in the emergence of Russell's logicist ideas should be more carefully explored.

In good measure, the fortune of Dedekind's work reflects the stormy development of logic and foundations during those days. Exaggeratedly abstract for most of his contemporaries, its abandonment of intuition for the sake of deductive rigor was even 'horrendous' ('*grässlich*') to Paul du Bois-Reymond (as recorded by young Hilbert in a visit in 1888 to Berlin; see [Dugac, 1976, 203 and 93]). In a letter to Klein, Dedekind himself made jokes about the negative reception that he expected [Dugac, 1976, 189].

By contrast, within a few years the book would be criticized for presenting things too quickly and easily, and for not making explicit the elementary logical basis of the whole theory (see, for example, [Frege, 1893, 1–3]; and also Russell (§61) and other logicians of the next generation). Praised by a few—especially Frege, Peirce, and above all Schröder and Hilbert—for having established a deep, hidden connection between arithmetic and logic, and indeed for demonstrating the logicist thesis, soon it would be severely criticized in the wake of the discovery of logical and set-theoretical paradoxes (compare [van Heijenoort, 1967] *passim*). At any rate, Dedekind's approach to the natural numbers, stripped from the logicist reading he made of it, fared much better than others. There was no problem to absorb it within axiomatic set theory, as happened already in Zermelo's key paper of 1908.

A surprising feature of *WSZ* is the total lack of references to Cantor's epoch-making work on transfinite set theory. In this writer's opinion, that was completely intentional and due to strained relations between both mathematicians. (Immediately after publication of *WSZ*, Cantor wrote to him defending some 'priority rights': the letter has not been preserved, but the issue is mentioned in Dedekind's letter to Weber of 24 January, 1888, in *Werke*, vol. 3, 489.) In fact, Cantor and his work was present in absence several times throughout the work. We have already commented on the most noteworthy instance, the case of art. 63 and the Cantor–Bernstein theorem. Another very striking instance comes

out in art. 161, which contains the definition of cardinal numbers. A footnote here indicates that Dedekind will restrict his definition to finite cardinals 'for reasons of simplicity and clarity', which obviously suggests that it would not be difficult to apply his treatment to more general cases. This can only mean a definition of transfinite cardinals based on the corresponding ordinals. So Dedekind is implicitly referring to Cantor's work (especially [Cantor, 1883], upon which see §46) and to the possibility of reworking it in analogy with his approach to the natural numbers. This would be done many years later in the context of axiomatic set theory.

A third instance of Cantor's presence in absence can be found by comparing the preserved 1887 draft and the final text. Dedekind eliminated a few words from his draft of the preface: having criticized those who take for simple complicated concepts like that of cardinality, he had written '(in opposition to Cantor)' (Cod. Ms. Dedekind III, 1, III: 'Gegensatz zu Cantor'). The reason why he chose to avoid this reference is unknown, but probably he wanted to avoid quarrels and disputes (this desire is explicit in the 1888 letter to Weber, mentioned above).

The impact of Dedekind's theory of sets, mappings and natural numbers in 20th-century axiomatic set theory deserves special mention. It began with the path-breaking paper [1908] by Zermelo. As is well known, Zermelo's axiomatization was intimately tied with his defence and reworking of the well-ordering theorem [Moore, 1982]. In order to produce a proof of well-ordering as simple and direct as possible, avoiding advanced concepts in set theory, Zermelo drew on the theory of chains and generalized it to deal with the transfinite case. His axiomatic system was based on a careful analysis of the work of Cantor and Dedekind, and he remarked that the axiom of infinity was 'essentially due to Dedekind' [Zermelo, 1908, 204]. Indeed, what Zermelo does here is to postulate the existence of a simply infinite set in the sense of Dedekind.

Chain theory was also employed by Thoralf Skolem in his new proof (1920) of Leopold Löwenheim's famous satisfiability result. Precisely the use of chains made it possible to fill a serious gap in Löwenheim's original proof, published five years earlier. The theory was employed again in 1922 by Casimiercz Kuratowski, in order to show how to eliminate the use of transfinite numbers, which by then was customary in mainstream mathematics, while these numbers had not yet been incorporated into axiomatic set theory! [Kuratowski, 1922, 76–108]. Generally speaking, the axiomatic treatment of the (finite) ordinals and of transfinite induction is closely related to the work of Dedekind.

BIBLIOGRAPHY

Cantor, G. 1883. *Grundlagen einer allgemeinen Mannichfaltigkeitslehre*, Leipzig: Teubner. [Repr. (without preface) in *Mathematische Annalen, 21,* 545–591. Also in *Gesammelte Abhandlungen* (ed. E. Zermelo), Berlin: Springer, 1932, 165–208. See §46.]

Dedekind, R. 1932. *Gesammelte mathematische Werke*, vol. 3 (eds. R. Fricke, E. Noether and Ö. Ore), Braunschweig: Vieweg. [Repr. New York: Chelsea, 1969.]

Dirichlet, P.G. Lejeune and Dedekind, R. 1879. *Vorlesungen über Zahlentheorie*, 3rd ed., Braunschweig: Vieweg. [Supplement XI in [Dedekind, 1932], 297–314. See §37.]

Dirichlet, P.G. Lejeune- and Dedekind, R. 1894. *Vorlesungen über Zahlentheorie*, 4th ed., Braunschweig: Vieweg. [Repr. New York: Chelsea, 1968.]

Dugac, P. 1976. *Richard Dedekind et les fondements des mathématiques* (*avec des nombreux textes inédits*), Paris: Vrin.

Ferreirós, J. 1993. 'On the relations between Georg Cantor and Richard Dedekind', *Historia mathematica, 20*, 343–363.

Ferreirós, J. 1999. *Labyrinth of thought. A history of set theory and its role in modern mathematics*, Basel and Boston: Birkhäuser.

Frege, G. 1893. *Grundgesetze der Arithmetik*, vol. 1, Jena: Pohle. [Repr. Hildesheim: Olms, 1966.]

Garciadiego, A. 1992. *Bertrand Russell and the origins of the set-theoretic 'paradoxes'*, Basel and Boston: Birkhäuser.

Grassmann, H. 1861. *Lehrbuch der Arithmetik für höhere Lehranstalten*, Berlin: Enslin.

Hallett, M. and Majer, U. (eds.) 2004. *David Hilbert's lectures on the foundations of geometry, 1891–1902*, Berlin: Springer.

van Heijenoort, J. (ed.) 1967. *From Frege to Gödel. A source book in mathematical logic, 1879–1931*, Cambridge and London: Harvard University Press.

Hilbert. D. 1905. 'Über die Grundlagen der Logik und Arithmetik', in *Verhandlungen des dritten internationalen Mathematiker-Kongresses in Heidelberg*, Leipzig: Teubner, 174–185. [References to English trans. in [van Heijenoort, 1967], 129–138.]

Hilbert, D. 1922. Neubegründung der Mathematik, *Abhandlungen der mathematischen Seminar Universität Hamburg*, 1, 157–177. [References to English trans. in W. Ewald (ed.), *From Kant to Hilbert*: *A source book in the foundations of mathematics*, New York: Oxford University Press, 1996, vol. 2, 1115–1134.]

Kuratowski, C. 1922. 'Une méthode d'elimination des nombres transfinis des raisonnements mathématiques', *Fundamenta mathematicae, 3*, 76–108.

Lipschitz, R. 1986. *Briefwechsel*, Braunschweig: Vieweg.

Moore, G.H. 1982. *Zermelo's axiom of choice. Its origins, development and influence*, Berlin: Springer.

Peano, G. 1891. 'Sul concetto di numero', *Rivista di matematica, 1*, 87–102, 256–267. [Repr. in [1959], 80–109.]

Peano, G. 1898a. *Formulaire de mathématiques*, vol. II, art. 2: *Arithmétique*, Turin: Bocca, 1898. [Part in [1959], 215–231.]

Peano, G. 1898b. 'Sul §2 del Formulario, t. II: Aritmetica', *Rivista di matematica, 6*, 75–89. [Repr. in [1959], 232–248.]

Peano, G. 1959. *Opere scelte*, vol. 3 (ed. U. Cassina), Rome: Cremonese.

Scharlau, W. (ed.) 1981. *Richard Dedekind 1831/1981. Eine Würdigung zu seinem 150. Geburtstag*, Braunschweig and Wiesbaden: Vieweg.

Schröder, E. 1895. *Vorlesungen über die Algebra der Logik*, vol. 3, Leipzig: Teubner. [Repr. New York: Chelsea, 1966.]

Tait, W.W. 1996. 'Frege versus Cantor and Dedekind: On the concept of number', in M. Schirn (ed.), *Frege*: *importance and legacy*, Berlin: Walter de Gruyter, 70–113.

Weyl, H. 1918. *Das Kontinuum*: *Kritische Untersuchungen über die Grundlagen der Analysis*, Leipzig: Veit. [Repr. with other work, New York: Chelsea, no year. English trans.: *The continuum* (trans. S. Pollard and T. Bole), New York: Dover, 1994.]

Zermelo, E. 1908. 'Untersuchungen über die Grundlagen der Mengenlehre, I', *Mathematische Annalen, 65*, 261–281. [References to English trans. in [van Heijenoort, 1967], 199—215.]

CHAPTER 48

HENRI POINCARÉ, MEMOIR ON THE THREE-BODY PROBLEM (1890)

June Barrow-Green

Drawing on his work on the qualitative theory of differential equations, in this memoir Poincaré developed a theory of periodic solutions that opened up an entirely new way of thinking about dynamical problems. It is famous both for containing the first description of mathematical chaos and for providing the basis for his acclaimed *Les méthodes nouvelles de la mécanique céleste* (1892–1899).

First publication. 'Sur le problème des trois corps et les équations de la dynamique', *Acta mathematica, 13* (1890), 1–270.

Reprint. In *Œuvres de Henri Poincaré,* vol. 7, Paris: Gauthier–Villars, 1952, 262–479.

Related articles: Newton (§5), Laplace (§18), Lyapunov (§51), Birkhoff (§68).

1 INTRODUCTION

Henri Poincaré (1854–1912) was educated at the *École Polytechnique* and the *École Nationale Supérieure des Mines*, and received his doctorate from the University of Paris in 1879. In 1881, after two years in charge of the analysis course at the University of Caen, he moved to the University of Paris, where from 1886 he occupied successively the chair of Mathematical Physics and Probability, and the chair of Mathematical Astronomy and Celestial Mechanics. He wrote more than 30 books and almost 500 papers, the most important being on function theory, geometry, topology, differential equations, celestial mechanics, electromagnetic theory, and the foundations of science.

Poincaré rose to international prominence during the early 1880s with his discovery of Fuchsian functions. During the same period, motivated by an interest in some of the fundamental questions of mechanics, in particular the problem of the stability of the solar system, he began his pioneering research on the qualitative theory of differential equations, work that laid the foundations for the memoir on three-body problem (hereafter, '*TBP*').

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

Through his work on Fuchsian functions and on the theory of differential equations he was led to recognise the importance of the topology (or as it was then called, 'analysis situs') of manifolds; so in the 1890s he began to study the topology of manifolds as a subject in its own right, effectively creating the new field of algebraic topology. Meanwhile, he continued to work and publish on celestial mechanics, the three volumes on *Les méthodes nouvelles de la mécanique céleste* ('New methods of celestial mechanics'; hereafter, '*CM*'), appearing in 1892, 1893 and 1899. Central to his success, both in topology and celestial mechanics, was his remarkable capacity for geometric visualisation. In the early years of the 20th century he became known to a much wider audience through his books of essays on the philosophy of mathematics and science.

The three-body problem is one of the most celebrated problems in celestial mechanics [Gautier, 1817]. Like many good problems, it is easy to state: three particles move in space under their mutual gravitational attraction; given their initial conditions, determine their subsequent motion. As is often the case with such problems, its importance lies as much in the mathematical advances generated by attempts at its solution as in the actual problem itself, and since its formulation by Isaac Newton many leading mathematicians have been attracted to it. But of the numerous papers published—more than 800 appeared between the years 1750 and 1900—none has continued to excite more attention than *TBP*.

2 ORIGIN AND SIGNIFICANCE OF THE THREE-BODY PROBLEM

The three-body problem established its place within the mathematical canon on the publication of Newton's *Principia* in 1687 (§5.10). From then on it became important to verify whether Newton's law of gravitation was capable of rendering a complete understanding of how celestial bodies move in three-dimensional space. This involved determining the relative motion of n bodies attracting one another according to the law. Newton himself had solved the two-body problem and so the three-body problem became the focus for attack.

Although Newton had been able to use geometry to solve the two-body problem, it rapidly became clear that the three-body problem required an analytical approach. Since each of the three particles has three position components and three velocity components, the problem is a system of order 18, and it can be represented by a system of nine second-order differential equations. But if these equations are to be solved exactly then 18 integrals of motion need to be found. By the end of the 18th century, largely as a result of the work of Leonhard Euler (1707–1783) and Joseph Louis Lagrange (1736–1813), 11 integrals had been discovered, and in 1843 Carl Jacobi (1804–1851) found a 12th. Jacobi was also responsible, together with Sir William Rowan Hamilton (1805–1865), for developing new methods for integrating the differential equations of a general dynamical system that turned out to be particularly useful in the context of the problem (compare §40.5).

Towards the end of the 1870s the American mathematical astronomer George William Hill (1838–1914) published two papers on the lunar theory that had a profound influence on the development of celestial mechanics in general and the three-body problem in particular. The key to Hill's success lay in his treatment of periodic solutions—one of his papers included the first new periodic solutions of the three-body problem since Lagrange's discovery of special periodic solutions in 1772—and this aspect of his work had a profound influence on Poincaré.

As the 19th century wore on and the impossibility of finding an exact solution to the problem looked increasingly likely, mathematicians shifted from searching for integrals to improving the approximations that resulted from the solution of the differential equations being given as infinite series. This involved attempting to eliminate the secular terms in the expansion in order to try to confine it to series in which the time only occurs within the arguments of the periodic terms. The difficulty of the problem also led mathematicians to consider a special simplified case. In this case—which was originally formulated by Euler and later termed the 'restricted' three-body problem by Poincaré—two of the bodies, known as the primaries, revolve around their centre of mass in circular orbits under the influence of their mutual gravitational attraction and so form a two-body system in which their motion is known. A third body, generally known as the planetoid, assumed massless with respect to the other two, moves in the plane defined by the two revolving bodies and, while being gravitationally influenced by them, exerts no influence of its own. The problem is then to ascertain the motion of the third body. Apart from its simplifying characteristics, the restricted problem also provides a good approximation for real physical situations, such as the problem of determining the motion of the Moon around the Earth, given the presence of the Sun. In the context of *TBP*, the restricted problem is important since it is the formulation upon which Poincaré based most of his work.

Potential solvers were also attracted to the three-body problem because of its intimate link with the fundamental question of the stability of the solar system. That is, the question of whether the planetary system will always keep the same form as it has now, or whether eventually one of the planets will escape from the system or, perhaps worse, experience a collision. If the Sun and the planets are considered as point masses—which is a reasonable approximation given that they are all virtually spherical and that their dimensions are extremely small when compared with the distances between them—and if only gravitational forces are taken into account, i.e. all other forces such as solar winds or relativistic effects are ignored, then the solar system can be modelled as a ten-body problem. Over the centuries many mathematicians and astronomers have been drawn to the stability problem. Two of the most notable were Lagrange and Pierre-Simon Laplace (1749–1827) (§18.4); both made significant advances, with Laplace believing that he had actually proved stability. Poincaré retained a fascination for the stability problem throughout his life [Poincaré, 1898] and he made no secret of the fact that it was an important spur behind much of his work in *TBP*.

3 POINCARÉ'S WORK BEFORE *TBP*

Almost from the beginning of his career Poincaré had been concerned with the fundamental problems of celestial mechanics, and many of the papers that he published during the 1880s relate to his interest in the subject. Arguably the most important is his acclaimed four-part memoir on curves defined by differential equations [Poincaré, 1881, 1882, 1885, 1886a], in which he initiated the qualitative theory of differential equations in the real domain. These papers are full of new ideas, many of which form the basis for results in *TBP*. The three-body problem featured prominently, and Poincaré was quite clear about its motivating role. When he began this work, research was, in effect, centred on studying the local properties of a solution to a differential equation; his approach was radically different. He

looked beyond local analysis and brought a global perspective to the problem by undertaking a qualitative study of the function in the whole plane. His objective was to provide a geometric study of the solution curves of a first-order differential equation, and indeed it was his geometrical insight that became one of the hallmarks of his research on differential equations. What was new and important was his idea of thinking of the solutions in terms of curves rather than functions, and it was this which marked a departure from the work of his predecessors whose research had been dominated by power-series methods. Since Poincaré's interest in the qualitative theory of differential equations was driven in part by his interest in the question of the stability of the solar system, he recognised the importance of considering the global properties of real as opposed to complex solutions, the latter having been the focus of earlier investigators.

Poincaré also produced several papers in which he addressed either a particular aspect of the three-body problem or a connected problem of celestial mechanics. These contain his initial researches into periodic solutions and his early investigations into the convergence of trigonometric series used in celestial mechanics. There are also papers in which he developed ideas and techniques which he used in *TBP* but which were generated in a more general context, such as his thesis on first-order partial differential equations and the paper [Poincaré, 1886b] on asymptotic series.

4 THE PUBLICATION OF *TBP*

TBP was published in 1890, but its journey to press began some five years earlier. In 1885 notices appeared announcing a mathematics competition to celebrate the 60th birthday of King Oscar II of Sweden and Norway. The organiser of the competition and one of the judges was the Swedish mathematician Gösta Mittag-Leffler (1846–1927); the other judges were Karl Weierstrass (1815–1897) and Charles Hermite (1822–1901). Four questions were set, of which one, posed by Weierstrass, required a solution to the n-body problem. The question, which reflected Weierstrass's long-standing interest in the problem, asked for a solution under the particular conditions that no collisions occur.

For Poincaré the competition acted as a stimulus to synthesise many of the ideas on which he had been working for several years. He had been interested in the question of the stability of the solar system for some time and had been building up a battery of techniques with which to launch an offensive. Many of these techniques originated in his research on the qualitative theory of differential equations in which he had first discussed the idea of stability. For the competition he had intended to tackle the n-body problem by starting with the general three-body problem and then extending his results, but the inherent difficulties led him to focus his attention almost exclusively on the 'restricted problem'. Despite not solving the n-body problem, his work on the restricted problem was recognised as outstanding and in January 1889 he was awarded the prize. The following year *TBP* was published as the winning entry in the competition.

A combination of royal patronage and carefully planned public relations meant that the competition gained recognition stretching well beyond the world of mathematics. In the numerous obituary notices and commentaries on Poincaré's œuvre, not only is *TBP* singled out for particular acclaim but the point is often made that it was as a consequence of winning the Oscar prize that Poincaré's fame became so widespread.

However, this widely applauded memoir is in fact very different from the version that actually won the prize. In the introduction to the published memoir, Poincaré mentioned that he had revised the essay for publication, but gave no indication of the nature and extent of his alterations. However, the discovery of a printed copy of the original essay personally annotated by him reveals all the revisions; in particular, it shows that some of the principal results for which the memoir is best known today are nowhere to be found in the original essay. More importantly, it shows that these new results are not simply extensions of previously existing ones; rather, they derive from Poincaré's discovery of a significant error, which he had made only a few days before the essay was due to be published. As a result of this discovery he was forced to rewrite a substantial proportion of the essay, a process that considerably delayed the ultimate publication of the memoir. Although the existence of the error was known to some of his contemporaries, its seriousness has only been recognised more recently [Barrow-Green, 1997]. It is now known that it was only as a result of correcting the error that Poincaré made his discovery of chaos.

Poincaré was naturally distressed by the discovery of the error and even questioned whether his original essay was still worthy of the prize. Although there was no question of the prize being rescinded, he had to pay for the reprinting of the memoir, the cost of which was considerably more than the prize he had originally won.

5 THE CONTENT OF *TBP*

The contents of Poincaré's memoir are summarised in Table 1. He adopted an unprecedented qualitative approach to the problem and its intrinsic dynamics. By using qualitative methods and focusing on how solutions behave rather than using quantitative methods and trying to find explicit formulae, he brought about a fundamental change in the way mathematicians thought about the problem.

Core to *TBP* is Poincaré's theory of periodic solutions. Having studied Hill's papers, he had seen the advantage of using periodic solutions to deal with problems of a general dynamical type, and in *TBP* he exploited this advantage to the full. In the first Part he discussed the underlying theory both from an analytical and from a geometric perspective. As a result the error occurs in both the geometry and the analysis. Its full implications became clear in the second Part of the memoir when he dealt with the application of the theory to systems with two degrees of freedom, in particular the restricted three-body problem.

In Poincaré's formulation of the restricted problem the position of the planetoid in phase space was described by two linear and two angular variables, x_1 and x_2, and y_1 and y_2 respectively, the latter taking the period 2π. The variables were connected by the integral of conservation of energy (art. 15):

$$F(x_1, x_2, y_1, y_2) = C. \tag{1}$$

He put the differential equations into Hamiltonian form

$$\frac{dx_i}{dt} = \frac{\partial F}{\partial y_i}, \quad \frac{dy_i}{dt} = -\frac{\partial F}{\partial x_i}, \quad i = 1, 2, \tag{2}$$

Table 1. Summary by Sections of Poincaré's memoir.

Section Arts; pp.	**'Title'; other included topics**
–; 3	'Introduction'.
1.1	'General properties of differential equations'.
1–4; 38	Notations and definitions. Method of majorants; application to partial differential equations. Integration of linear equations with periodic coefficients.
1.2	'Theory of invariant integrals'.
5–8; 42	Properties of the equations of dynamics. Definition, transformation and use of invariant integrals. Recurrence theorem.
1.3	'Theory of periodic solutions'.
9–14; 88	Existence of periodic solutions; characteristic exponents. Periodic solutions of the equations of dynamics. Calculation of characteristic exponents. Asymptotic solutions, including of the equations of dynamics.
2.1	'The case with two degrees of freedom'.
15; 15	Geometric representations.
2.2	'Study of asymptotic surfaces'.
16–19; 47	Statement of the problem. First, second, third approximation.
2.3	'Miscellaneous results'.
20–22; 38	Periodic solutions of the second class. Divergence of Lindstedt's series. Non-existence of single-valued integrals: denseness of periodic solutions.
2.4	'Attempts at generalisation'.
23; 5	The n-body problem.

which, in accordance with the qualitative theory that he had previously developed, he regarded as defining flows on a three-dimensional surface. His brilliant insight was to recognise that rather than considering the flow in the entire three-dimensional space, it was much more convenient to consider the first return map induced by the flow on a two-dimensional surface of section S transverse to the flow (art. 8). (Today such surfaces of section are known as 'Poincaré sections'.) This map is defined by choosing a point M on S at which S is intersected by a flow line; then the image of M under the map is the point M' at which that flow line first intersects S again. Thus in the three-dimensional space a periodic solution corresponds to a closed curve, but under the map a 2π-periodic solution corresponds to a fixed point and a $2\pi k$-periodic solution corresponds to a cycle of period k.

To form the power series expansions of the solutions to the equations Poincaré used the mass of the smaller of the primaries, μ, as the parameter. This is because when $\mu = 0$ the problem reduces to the Kepler problem, that is, attraction by a single fixed centre, and he could employ the strategy of starting with a particular solution for which $\mu = 0$ before varying μ analytically to see if solutions existed for very small values of μ.

Poincaré developed the Hamiltonian F in powers of μ,

$$F = F_0 + \mu F_1 + \mu^2 F_2 + \cdots, \tag{3}$$

where F_0 depends only on x and $F_1, F_2, \ldots$ are periodic functions of period 2π with respect to y (art. 16). He supposed that when $\mu = 0$ there existed periodic solutions of the form

$$x_i = \phi_i(t), \qquad y_i = \psi_i(t). \tag{4}$$

He then showed that there also existed periodic solutions of the form

$$x_i = \phi_i(t) + \xi_i, \qquad y_i = \psi_i(t) + \eta_i, \tag{5}$$

where

$$\xi_i = S_i \exp(\alpha_k t), \qquad \eta_i = T_i \exp(\alpha_k t). \tag{6}$$

The S_i and T_i were periodic functions of t, and the α_k are certain constants which he called 'characteristic exponents' (art. 10). Importantly, he realised that it was the form of the characteristic exponents that indicated the stability of the solutions. If the characteristic exponents are imaginary then the periodic solution is stable, otherwise it is unstable. In his discussions on stability Poincaré used the definition proposed by Siméon-Denis Poisson (1781–1840), that the motion of a point is regarded as stable if it returned infinitely often to positions arbitrarily close to its initial position.

As Poincaré recognised, one of the great advantages of periodic solutions is that they provide a natural starting point for studying and classifying other nearby solutions. And it was by studying solutions only slightly differing from a given periodic solution that he was led to his remarkable discovery of asymptotic solutions: solutions which either slowly approach or slowly move away from an unstable periodic solution (art. 13). He showed that in the three-dimensional solution space of the restricted problem, these asymptotic solutions generate families of curves which fill out surfaces and which asymptotically approach the curve representing the generating unstable periodic solution, and that these surfaces correspond to curves in the transverse section. In order to gain an understanding of the behaviour of these asymptotic solutions, Poincaré investigated the nature of the curves on the transverse section. This investigation required what was to be another important topic in *TBP*, and one particularly significant with regard to the error: the theory of invariant integrals (art. 6–8).

Although Poincaré was not the first to recognise the existence and value of invariant integrals—they are earlier encountered in the work of Joseph Liouville (1809–1882) and Ludwig Boltzmann (1844–1906)—he was responsible for developing the general theory that revealed that the existence of an invariant integral is a fundamental property of Hamiltonian systems of differential equations. His theory of invariant integrals also led him to his renowned recurrence theorem: that given a region of phase space, however small, there will be trajectories which traverse it infinitely often (art. 8). In other words, at some future time the system will return arbitrarily close to its initial situation and will do so infinitely often.

Poincaré concluded his discussion of invariant integrals with a series of theorems characterised by their geometric nature (art. 8), and in the original version of *TBP* it was in one of these theorems that the fundamental error in his geometry had occurred. He had failed to take proper account of the exact geometric nature of a particular curve. He thought that he had proved a particular curve was closed when, as he later realised, it was actually self-intersecting. In essence he had failed to take into account the full range of possibilities consistent with the constraints imposed by the existence of the invariant integral.

The error was reflected in Poincaré's analytical description of the asymptotic solutions. When he had originally calculated the series expansions for these solutions, he had assumed that the series were convergent. But, as he subsequently discovered, the series were actually divergent (art. 13); they belonged to the class of series now known as asymptotic and for which he himself had provided the first formal definition [Poincaré, 1886b].

In Poincaré's geometric representation of the restricted problem, a generating unstable periodic solution and its accompanying family of asymptotic solutions are represented in the three-dimensional solution space by a closed curve and two asymptotic surfaces (art. 16). To understand the behaviour of the asymptotic solutions, he sought the exact equations (in infinite series expanded in powers of the parameter μ) for the asymptotic surfaces. He considered the intersections of the surfaces with a transverse section and proceeded by successive stages of approximation (arts. 17–19).

In his original analysis, Poincaré was led to the mistaken result that the intersections of the asymptotic surfaces with the transverse section were represented by closed curves and hence that the asymptotic surfaces were closed. Furthermore, inherent in this result was the implication of stability in the sense that the solutions remained confined to a given region of space. In other words, he believed that he had proved that for sufficiently small values of the parameter μ there was, relative to a given unstable periodic solution, a set of asymptotic solutions which could be considered stable, that these solutions were well behaved and that they could be completely understood.

In his revised analysis, Poincaré first proved that the series for the asymptotic solutions were not convergent. He then established that the curves representing the asymptotic surfaces were not closed but self-intersecting and, moreover, they intersected infinitely often (art. 19). He called the trajectories that passed through the point of intersection 'doubly asymptotic'. (Later, in *CM*, he called them 'homoclinic solutions', and the points of intersection are now known as 'homoclinic points'.) Poincaré's description in *TBP* of doubly asymptotic trajectories is the first mathematical description of chaotic motion in a dynamical system. Although he drew little attention to the behaviour that he had discovered and made no attempt to draw a diagram, he was profoundly disturbed by his discovery, and almost a decade elapsed before he published anything further on the subject.

The penultimate chapter of *TBP* contains supplementary results connected with the three-body problem. The first of these is a proof of the existence of periodic solutions making more than one revolution around the origin (art. 20). This is notable for including Poincaré's conjecture concerning the denseness of the periodic solutions: that given any particular solution to the restricted problem it should be possible to find a periodic solution (which may have an extremely long period) such that the difference between the two solutions is as small as desired for any given length of time, providing no escape or collision occurs. Poincaré did not prove the conjecture himself but it was later shown to be true.

The chapter also contains a proof of the non-existence of any new integral of the restricted problem (art. 22) that was an important complement to a result published by Heinrich Bruns (1848–1919) in 1887 which showed that no new algebraic integral of the general three-body problem could exist. Also important is Poincaré's discussion of the purely trigonometric series known as Lindstedt's series (art. 21). This series, which contains no secular terms, was named after the Swedish astronomer Anders Lindstedt (1854–1939) who had sought to show that it could be used to solve a particular form of second-order differential equation. Poincaré demonstrated, contrary to what had previously been thought, that the series were not uniformly convergent for all the values of the arbitrary constants of integration they contain, although his discussion was incomplete as he gave no consideration to the circumstances under which convergence could occur.

In the final chapter (art. 23), Poincaré provided a brief discussion on the difficulties in generalising his earlier results to the n-body problem. He showed, for example, how, in a particular case, an increase from two to three in the number of degrees of freedom led either to the problem of 'small divisors' or to inscrutable integrals. Either way the problem was intractable.

6 POINCARÉ ON CELESTIAL MECHANICS AFTER *TBP*

Poincaré's book *CM* (1892–1899) contains the principal ideas from *TBP* but in a more fully explained and developed form. A greater number of applications of the theory are included, as well as a substantial amount of new material, with the attention being as much on the general three-body problem as on the restricted problem. The first volume includes an amplified treatment of periodic solutions, asymptotic solutions and the non-existence of new uniform integrals. The second volume is devoted to the perturbation methods of mathematical astronomers; while the third volume, which is characterised by Poincaré's geometrical ideas, contains a discussion of invariant integrals and stability. In the third volume he returned for the first time to the subject of doubly asymptotic solutions, further developing the theory and discovering a second more complex type of solution. These new solutions, which he called 'heteroclinic solutions', are associated with two unstable periodic solutions rather than one, and are correspondingly more complicated.

Poincaré also published several short papers of a general nature on the three-body problem and on the stability of the solar system. These papers embrace a greater practical perspective than *TBP* and were a well judged response to the need for a more popular presentation of his ideas. They include a synopsis of the memoir specifically designed to be accessible to astronomers and those whose interest in the three-body problem was motivated by practical considerations, as well as an almost completely descriptive exposition of results relating to the restricted three-body problem [Poincaré, 1891a, 1891b].

7 THE RECEPTION OF *TBP*

As the winning entry in the Oscar competition *TBP* attracted considerable attention, and it met with an enthusiastic reception. There was, however, one detractor, the astronomer Hugo Gyldén (1841–1896) who mistakenly believed that he had already discovered similar

results. But despite the positive response, it is conspicuous that almost all of the reviews of the memoir failed to include any discussion of Poincaré's doubly asymptotic solutions; typical is the detailed account of *TBP* by Edmund Taylor Whittaker (1873–1956) which appeared as part of his BAAS report on the three-body problem [Whittaker, 1899]. An exception was provided by Hermann Minkowski (1864–1909) who, in his review for the *Jahrbuch*, openly acknowledged the difficulties associated with the solutions [Minkowski, 1893].

However, despite its warm reception, *TBP* did create problems for many of its readers. Not only was it fiercely demanding and full of new mathematics; the difficulties were compounded by Poincaré's customary lack of detail. Astronomers in particular struggled with it: even Hill was led to criticise publicly some of the results. Although, as Poincaré himself showed, little of Hill's criticisms stood up to rigorous scrutiny, they reveal just where many of the problems lay.

In 1891 some interesting observations on particular results in *TBP* were made by Lord Kelvin (1824–1907) who, having had the memoir brought to his attention by Arthur Cayley (1821–1895), was especially struck by the relationship between some of Poincaré's results and some conclusions of his own. In particular, he drew attention to the similarity between Poincaré's conjecture concerning the denseness of the periodic solutions and a proposition of Maxwell concerning the distribution of energy.

One of the first of Poincaré's ideas from *TBP* to emerge in a different context was that of his recurrence theorem. The theorem appeared to demonstrate the futility of contemporary efforts to deduce the second law of thermodynamics from classical mechanics, and in 1896 a debate took place between Ernst Zermelo (1871–1953), who believed that Poincaré's theorem disproved the absolute validity of the second law, and Boltzmann, who believed in the correctness of Poincaré's theorem but disputed Zermelo's application of it. Although Zermelo and Boltzmann's debate came to an end within a year, the controversy continued to arouse interest and eventually became one of the sources for the foundation of modern ergodic theory.

8 THE RESOLUTION OF THE THREE-BODY PROBLEM AND SOME LATER DEVELOPMENTS

With the publication of *TBP* work on the three-body problem intensified. The specification of the problem in the Oscar competition had included the assumption that no collisions between the bodies would take place and Poincaré had based his analysis accordingly. But if a complete solution to the problem was to be found then collisions had to be taken into account. Since collisions are described by singularities in the differential equations, this raised questions of regularisation. Could the singularities be removed by a change of variable so that the motion could be followed through the point of collision? Could singularities other than collision singularities exist?

In 1896 Paul Painlevé (1863–1933) proved not only that the only singularities are collisions but also that a mathematical solution to the three-body problem could be found providing it was possible to define precisely the initial conditions corresponding to a collision. The person who found that solution was Karl Sundman (1873–1959), an astronomer

at the Helsinki Observatory. In 1907 Sundman completely defined the initial conditions for both binary and triple collisions. Not only were Sundman's results quite remarkable: the methods he used were surprisingly simple. Essentially they depended on the application of an extension to a well-known theorem of Augustin Louis Cauchy (1789–1857) on the existence of solutions to differential equations (§25.5).

Although the significance of Sundman's achievement was recognised by his contemporaries, within about a decade it was almost completely forgotten. This neglect can be partly attributed to the practical limitations of his results. The rate of convergence of the series which he had derived was extremely slow and so for practical purposes the classical divergent series were thought to be more useful. In addition, the results that Sundman obtained furnished no qualitative information about the nature of the motion. He had provided a mathematical solution but not one which revealed general information about the form of the trajectories; hence it left unresolved many issues surrounding the problem.

An important complement to *TBP* was provided by Alexander Lyapunov (1857–1918) in his qualitative investigation of 1893 into the theory of the stability of motion (§51). The subject of stability was also taken up by Tullio Levi-Civita (1873–1941), who drew upon both Poincaré's and Lyapunov's ideas. *TBP* was also a signal influence on the work of Jacques Hadamard (1865–1963) on the theory of geodesics on surfaces of positive and of negative curvature, of Ivar Bendixson (1861–1935) on ordinary differential equations, and of Elie Cartan (1869–1951) on the theory of invariant integrals.

But the mathematician most influenced by *TBP* and *CM* was undoubtedly George Birkhoff (1884–1944). Incorporating a vigorous use of topology, Birkhoff both generalised and extended Poincaré's ideas; and, like Poincaré, he made the periodic motions play a central role in his theory. Birkhoff's deep study of Poincaré's work is evident from his first publication devoted to theoretical dynamics in which he introduced the idea of 'recurrent motion' as a natural extension of periodic motion [Birkhoff, 1912]. But the result which ties him irrevocably with Poincaré, and the one for which he is arguably most famous, is his resolution of Poincaré's last geometric theorem [Birkhoff, 1913] in which he confirmed the existence of an infinite number of periodic solutions for the restricted three-body problem for all values of the mass parameter μ. The essential ideas of this paper, together with many other ideas derived from Poincaré, can be found in Birkhoff's acclaimed book *Dynamical systems* [Birkhoff, 1927] (§68).

In addition, the question of the convergence of Lindstedt's series provided the starting point for some remarkable 20th-century developments. Poincaré had shown that, apart from some exceptional cases, the series were divergent. There was, however, one proviso. He had made it clear that he had not given a rigorous proof for the cases when the frequencies can be fixed in advance. With the work of A.N. Kolmogorov, V.A. Arnold and J. Moser, which began in the 1950s, it is now known that in these latter cases the majority of the formal series solutions are in fact convergent. Their results form the basis for what is now known as Kolmogorov–Arnold–Moser (KAM) theory which provides methods for integrating perturbed Hamiltonian systems valid for infinite periods of time. Of particular significance is the fact that KAM theory conclusively establishes the existence of convergent series solutions for the n-body problem.

Despite the success of *TBP* it is conspicuous that during the early years of the 20th century no serious attempt was made to investigate further the behaviour of Poincaré's

asymptotic solutions. This can largely be explained by the inability to undertake a quantitative analysis due to lack of computing power. But with the arrival of the modern digital computer such an analysis has now become possible, with the result that the last 30 years has seen an explosion of research into non-linear systems with the concomitant unfolding of the mathematical theory of chaos.

Finally, there is another reason why Poincaré's asymptotic solutions may have been ignored: their apparently random behaviour did not fit in with the then widely accepted Laplacian model of a clockwork universe. Indeed, it may be that this belief in some kind of ultimate order was partly responsible for Poincaré himself missing the chaotic behaviour in his original essay.

BIBLIOGRAPHY

Barrow-Green, J. 1997. *Poincaré and the three body problem*, Providence: American Mathematical Society; London: London Mathematical Society.

Birkhoff, G.D. 1912. 'Quelques théoremes sur le mouvement des systèmes dynamiques', *Bulletin de la Société Mathématique de France, 40*, 305–323. [Repr. in *Collected mathematical papers*, vol. 1, 654–672.]

Birkhoff, G.D. 1913. 'Proof of Poincaré's geometric theorem', *Transactions of the American Mathematical Society, 14*, 14–22. [Repr. in *Collected mathematical papers*, vol. 1, 673–681.]

Birkhoff, G.D. 1927. *Dynamical systems*, Providence: American Mathematical Society. [See §68.]

Diacu, F. and Holmes, P. 1996. *Celestial encounters*, Princeton: Princeton University Press.

Gautier, A. 1817. *Essai historique sur le problème des trois corps*, Paris: Courcier.

Minkowski, H. 1893. Review of *TBP, Jahrbuch über die Fortschritte der Mathematik, 22* (1890), 907–914.

Poincaré, H. *Works. Œuvres*, 11 vols., Paris: Gauthier Villars, 1916–1956.

Poincaré, H. 1881, 1882. 'Mémoire sur les courbes définies par une équation différentielle', *Journal de mathématiques pures et appliquées, (3) 7*, 375–422; *(3) 8*, 251–196. [Repr. in *Works*, vol. 1, 3–44; 44–84.]

Poincaré, H. 1885, 1886a. 'Sur les courbes définies par les équations différentielles', *Journal de mathématiques pures et appliquées, (4) 1*, 167–244; *(4) 2*, 151–217. [Repr. in *Works*, vol. 1, 90–161; 167–222.]

Poincaré, H. 1886b. 'Sur les integrals irrégulières des équations linéaires', *Acta mathematica, 8*, 295–344. [Repr. in *Works*, vol. 1, 290–332.]

Poincaré, H. 1891a. 'Sur le problème des trois corps', *Bulletin astronomique, 8*, 12–24. [Repr. in *Works*, vol. 7, 480–490.]

Poincaré, H. 1891b. 'Le problème des trois corps', *Revue générale des sciences pures et appliquées, 2*, 1–5. [Repr. in *Works*, vol. 7, 529–537.]

Poincaré, H. 1898. 'On the stability of the solar system', *Nature, 58*, 183–184. [Repr. in *Works*, vol. 8, 538–547.]

Whittaker, E.T. 1899. 'Report on the progress of the solution of the problem of three bodies', *BAAS Report*, 121–159.

CHAPTER 49

OLIVER HEAVISIDE, *ELECTRICAL PAPERS* (1892)

Ido Yavetz

In this book Heaviside brought together his main writings on circuit theory and the inductive properties of wires, and his thoughts on electromagnetism. He also extended operator and vector algebras.

First publication. 2 volumes, London: Macmillan, 1892. 560 + 587 pages.

Photoreprint. New York: Chelsea Publishing Company, 1970 [with correction of errata].

Related articles: Hamilton (§35), Thomson and Tait (§40), Maxwell (§44).

> 'Familiarity with the working out of physical problems breeds contempt for the idea of requiring a special demonstration of the possibility of what seems to be necessary' (vol. 1, 148).

1 GENERAL OUTLINE OF THE *ELECTRICAL PAPERS*

Oliver Heaviside was born in London in 1850, and died in Torquay in 1925. His formal schooling ended when he was 15 years old. By his own account he did not obtain much scientific knowledge from it—arithmetic, a smattering of trigonometry, and an introduction to Euclidian geometry that he abhorred to the end of his days. He was, however, a voracious reader, and spent considerable time in public libraries. There, among other things, he made his first acquaintance with the work of J.C. Maxwell (1831–1879). At the age of sixteen, he obtained a telegraph operator's job, with the help of Sir Charles Wheatstone—his uncle on his mother's side. His interest in the principles of telegraphy became the motivation behind much of his scientific work. Heaviside never attended any institution of higher learning. His outstanding achievements in mathematics, physics, and engineering science were the fruits of self-education in a professional age that made such an endeavor nearly impossible. For more details on Heaviside's biography and eccentric character, see [Nahin, 1988] and [Yavetz, 1995, ch. 1].

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

The two volumes of Heaviside's *Electrical papers* contain papers that he wrote between 1872 and 1892. They encompass the results of his most creative scientific years, and also reflect his remarkable process of self-education. What the title does not reflect is that the two volumes possess a far greater degree of formal cohesion and continuity of subject matter than one might expect from a collection of scientific papers. As he pointed out in the preface to the first volume (p. vi):

> [. . .] it had been represented to me that I should rather boil the matter down to a connected treatise than republish in the form of detached papers. But a careful examination and consideration of the material showed that it already possessed, on the whole, sufficient continuity of subject-matter and treatment, and even regularity of notation, to justify its presentation in the original form. For, instead of being, like most scientific reprints, a collection of short papers on various subjects, having little coherence from the treatise point of view, my material was all upon one subject (though with many branches), and consisted mostly of long articles, professedly written in a connected manner, with uniformity of ideas and notation. And there was so much comparatively elementary matter (especially in what has made the first volume) that the work might be regarded not merely as a collection of papers for reference purposes, but also as an education work for students of theoretical electricity.

The *Electrical papers* offer an advanced exposition, as well as many novel contributions to two basic themes: the theory of electromagnetic field dynamics due to Maxwell (§44), and an extension of linear circuit theory to the case of continuous transmission lines. The following pages focus upon Heaviside's mathematical innovations, namely, his contributions to the formulation of vector algebra, and his controversial version of the operational calculus. However, he developed the first of these topics in close connection to his study of field electrodynamics, and the second in equally intimate connection to his circuit analysis. His mathematical thinking was always closely guided by the physics and engineering problems that he engaged; the *Electrical papers* reflect the growth of his mathematical ideas in association with their applications.

Generally, the main chapter headings in the two volumes are the original titles of the articles that Heaviside published in various periodicals. Many of these articles stretched into series—some of them hundreds of pages in length. These series were originally published in separate sections, which were also retained when republished in the two volumes of *Electrical papers*.

The articles in volume I may be grouped as follows. Arts. 1–23, pp. 1–195, were first written between 1872 and 1883. One article containing a particularly interesting discussion of the propagation of currents in transmission lines, was written in 1882 but published for the first time here (art. 20, pp. 141–179). The others were published (number of articles in parentheses) in *English mechanic* (1), the *Telegraphic journal* (2), the *Philosophical magazine* (10), *The electrician* (4), and *The Journal of the Society of Telegraph Engineers* (5). They contain discussions of various problems pertaining to telegraphy. Some, mostly the earlier papers, provide circuit-theoretical analysis of end-instruments in telegraphic systems, namely, transmitters, receivers, and batteries. The later papers in this group concentrate mainly on the propagation of signals along telegraph lines.

Arts. 24–30, pp. 195–560. Of these, Art. 29 (pp. 416–428), published in March 1885 in *The Journal of the Society of Telegraph Engineers*, is exceptional: Heaviside's contribution to an on-going debate on 'The seat of the electromotive forces in the Voltaic cell', quoting the title of an address given by Oliver Lodge—the main protagonist in this debate—at the 1884 Montreal meeting of the British Association for the Advancement of Science [Lodge, 1884]. The next six articles reflect a shift both in his publication style and subject matter: an extensive survey and re-formulation of the elements of Maxwell's electromagnetic field dynamics, and only very short, incidental comments on circuits and transmission lines. Taken together, they constitute a Maxwellian treatise running well over 300 pages. In this context Heaviside began to develop vector algebra, as what he considered to be the proper mathematical language for the discussion of three-dimensional force fields.

Volume II is considerably less organized and more difficult to describe. While writing the papers that were later collected in it, Heaviside had become embroiled in a very bitter dispute with William Henry Preece, the chief engineer of the British Post Office. Preece succeeded in pressuring the editor of *The electrician* into discontinuing Heaviside's on-going series 'On electromagnetic induction and its propagation'. He transferred material that he intended for this series into another series, 'On the self-induction of wires', that he had been publishing at the same time in the *Philosophical magazine*. This series was prematurely terminated also, owing partially, perhaps, to further disruptive efforts by Preece. The ban did not last long, and Heaviside managed to publish nearly all the material in new papers, so that very little was actually suppressed. However, the two discontinued series had been carefully planned to present a double-ended exposition of Maxwell's theory and transmission line analysis. In one series Heaviside planned to show how under various restrictions Maxwell's equations reduce to characteristic circuit equations; in the other series he planned to show how circuit laws may be generalized into Maxwell's field equations. As he explained (vol. 2, 76):

> In another place (Phil Mag., Aug., 1886 and later) the method adopted by me [. . .] was to work down from a system exactly fulfilling the conditions involved in Maxwell's scheme, to simpler systems nearly equivalent, but more easily worked. Remembering that Maxwell's is the only complete scheme in existence that will work, there is some advantage in this; also, we can see the degree of approximation when a change is made. In the following I adopt the reverse plan of rising from the first rough representation of fact up to the more complete. This plan has, of course, the advantage of greater intelligibility to those who have not studied Maxwell's scheme in its complete form; besides being, from an educational point of view, the more natural plan.

The premature discontinuation of the series, and the publication of their material in different frameworks, practically destroyed Heaviside's ambitious plans for publication. He never forgave the damage he suffered at Preece's hands, and never missed an opportunity to put in a bad word for him. The lesser degree of organization of Volume II is in large measure explained by these events.

The volume is dominated by three long series:

Art. 35, pp. 39–154, from *The electrician* (April 1886 – December 1887). Sections 25–31 are the direct continuation of 'On electromagnetic induction and it's propagation' from

Volume 1. The new publication scheme that was cut short by Preece begins with Section 32, and comes to an abrupt end in Section 47.

Art. 40, pp. 168–323, from *Philosophical magazine* (August 1886 – July 1887) 'On the self-induction of wires'. The second half of Heaviside's publication plan, only the first seven sections of this series had been published; the eighth first appeared here. Further material originally intended for this series was to appear in other papers.

Art. 43, pp. 375–467, from *Philosophical magazine* (February – December 1888) 'On electromagnetic waves, especially in relation to the vorticity of the impressed forces; and the forced vibrations of electromagnetic systems'. This six-part series contains most of the material that had been denied publication in the previous two. The style, however, is quite different, being very technical and mathematically more demanding. Here Heaviside solved electromagnetic equations by expressing them operationally, expanding the operational expressions in Bessel functions and Fourier series, and then transforming the expanded expressions back into terms of time and the space coordinates.

Two other articles are of special interest for the present discussion.

Art. 42, pp. 355–374, from *Philosophical magazine* (December 1887) 'On resistance and conductance operators, and their derivatives, inductance and permittance, especially in connection with electric and magnetic energy'. This paper was Heaviside's first concentrated attempt to outline the ideas that are now known as 'Heaviside's operational calculus'.

Art. 44, pp. 468–485, from *Philosophical magazine* (January 1889) on 'The general solution of Maxwell's electromagnetic equations in a homogeneous isotropic medium. Especially in regard to the derivation of special solutions, and the formulae for plane waves'. In physical subject matter, this article relates to arts. 31 and 43; mathematically, the style follows that of the latter. The solutions of the equations are written operationally, and then interpreted by expanding them into power series of the differential operator and its inverse. This procedure was later to become Heaviside's main approach to the operational calculus.

The remainder of the volume engages in various problems of circuit theory, electromagnetic fields and waves, and questions regarding units and nomenclature in electromagnetism. Particularly remarkable is art. 52, pp. 521–574, from *Philosophical transactions of the Royal Society* (1892) 'On the forces, stresses, and fluxes of energy in the electromagnetic field'. Over a mere 54 pages, Heaviside produced a systematic introduction to vector algebra, a general discussion on the dynamics of stresses and strains in elastic and dissipative media, and an advanced study of electromagnetic field dynamics. The presentation is highly condensed, notationally idiosyncratic, and the paper was considered unusually hard to follow by Heaviside's contemporaries, among them accomplished mathematical physicists like H.R. Hertz, George Francis FitzGerald (1851–1901), and Horace Lamb.

2 THE ALGEBRA OF VECTORS

The modern reader of Heaviside's *Electrical papers* should not have difficulty in dealing with Heaviside's introduction and use of vector algebra; for by and large it is the vector algebra that we now learn. From him we have adopted the convention of using bold letters to signify vectors. The dot and the diagonal cross that signify the scalar and vector products (respectively $\mathbf{A} \cdot \mathbf{B}$ and $\mathbf{A} \times \mathbf{B}$) are due to Willard Gibbs (1839–1903) (§35.5). Heaviside

did not like that, and preferred to write the scalar product of the vectors **A** and **B** simply as '**AB**'. The vector product he wrote, following W.R. Hamilton (1805–1865) and Peter Guthrie Tait (1831–1901), as 'V**AB**'. Another difference that does not really create too much of an inconvenience is Heaviside's general tendency to avoid the symbol ∇ for the vector operator $(d/dx, d/dy, d/dz)$ in equations that express physical ideas, and to employ instead the expressions div(**A**), conv(**A**), and curl(**A**) to denote what we usually write out as $\nabla \cdot \mathbf{A}$, $-\nabla \cdot \mathbf{A}$ and $\nabla \times \mathbf{A}$ respectively. The words 'div' and 'curl' are still widely associated with the first and third, and only 'convergence' is no longer current for negative divergence.

In some ways, the appearance of vector algebra is the natural and undramatic culmination of a very long history [Crowe, 1967]. The practice of breaking directed magnitudes like velocity into components is as old as the pseudo-Aristotelian 'Mechanical problems'. Here the parallelogram rule for the combination of directed magnitudes is explicitly and correctly spelled out [Aristotle, 1953, 381–387]. If this book was composed in the late 4th century B.C., as many experts estimate, then the historical roots of vector algebra are very old indeed. The remarkable thing about vector algebra as we currently know it, is that it did not emerge from a direct attempt to formalize such widely familiar applications. Instead, it made its first appearance in the context of Hamilton's highly innovative and idiosyncratic invention of quaternions (§35). Hamilton believed that quaternions are destined to become the universal language of mathematical physics. Tait agreed, and following in Hamilton's footsteps composed a concise textbook on quaternions ([Tait, 1867]: compare §35.4). In their quaternionic context, however, vectors possessed formal properties that proved awkward for application to physical problems. It was then left for Gibbs and Heaviside to extract the vector from its quaternionic foundations, and establish it on its own formal grounds. For Heaviside the need to do this became apparent in the course of reading Maxwell's theory of electromagnetic field dynamics. What Maxwell saw in Hamilton's quaternions, and how he employed them in his *Treatise on electricity and magnetism*, he expressed quite openly [1892, vol. 2, 257]:

> In this treatise we have endeavoured to avoid any process demanding from the reader a knowledge of the Calculus of Quaternions. At the same time we have not scrupled to introduce the idea of a vector when it was necessary to do so.

Among the various properties of Quaternions we note that in the relationship $ij = k$, i and j do not represent the same mathematical idea. Rather, the second index, j, represents the transformation of a scalar magnitude into a vector along the y-axis, while the first, i, rotates this vector by 90° about the x-axis, bringing it into alignment with the z-axis. As long as this peculiarity is kept in mind, it should be clear from the foregoing that $ijk = -1$, $ikj = 1$, and so on. Heaviside was justifiably annoyed by this double use. But it may be noted that no deep-rooted contradiction lies at the bottom of this apparent inconsistency; for it can be avoided by interpreting the quaternion $\boldsymbol{q}$ as $\sum_{n=0}^{3} q_n e_n$, where

$$e_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad e_1 = \begin{pmatrix} 0 & -i \\ -i & 0 \end{pmatrix}, \quad e_2 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \quad \text{and}$$

$$e_3 = \begin{pmatrix} -i & 0 \\ 0 & i \end{pmatrix}, \tag{1}$$

with i standing for $\sqrt{-1}$.

Following these observations, a quaternion $\boldsymbol{q}$ may be defined as a four-part entity, that consists of a scalar component, q_0, and three space components q_1, q_2, q_3. It may be written alternatively as

$$\boldsymbol{q} = (q_0, q_1, q_2, q_3) = q_0 + iq_1 + jq_2 + kq_3 = q_0 + \mathbf{q}, \tag{2}$$

where q_0 is a scalar and $\mathbf{q}$ a space vector.

The properties of quaternion multiplication may now be extracted in a straightforward manner from the definitions above, with the understanding that if a is a scalar quantity then $ia = ai$. It turns out that if $\boldsymbol{p}$ and $\boldsymbol{q}$ are quaternions, then

$$\begin{aligned} \boldsymbol{pq} = (p_0 + \mathbf{p})(q_0 + \mathbf{q}) = p_0q_0 + p_0\mathbf{q} + q_0\mathbf{p} - [p_1q_1 + p_2q_2 + p_3q_3] \\ + \big[i(p_2q_3 - p_3q_2) + j(p_3q_1 - p_1q_3) + k(p_1q_2 - p_2q_1)\big]. \end{aligned} \tag{3}$$

The modern reader will recognize the two expressions in square brackets as the scalar and vector products of standard vector algebra. Quaternion multiplication contains these two, and may appear, therefore, as a manner of unifying them under a single generalized operation. Note, however, that if the scalar component of quaternion $\boldsymbol{p}$ is 0, then by the foregoing rules of quaternion multiplication:

$$\boldsymbol{p}^2 = -\mathbf{p}^2. \tag{4}$$

This is all as should be, so long as quaternions are regarded as extending the concept of imaginary numbers. If, however, the vector component of the quaternion is meant to represent a physical entity, like fluid flow or an electromagnetic field, the fact that its square is negative represents an awkward feature, with no straightforward physical meaning in most cases. As Heaviside wrote in one of his earliest sketches of vector algebra (vol. 2, 3):

> It is a matter of great practical importance that the notation should be such as to harmonize with Cartesian formulae, so that we can pass from one to the other readily, as is often required in mixed investigations, without changing notation. This condition does not appear to me to be attained by Professor Tait's notation, with its numerous letter prefixes, and especially by the $-S$ before every scalar product, the negative sign being the cause of the greatest inconvenience in transitions.

In later work, Heaviside as well as Gibbs, used this to indicate that regardless of their pure mathematical merits, quaternions are ill-suited as means for expressing physical ideas mathematically (EMT, vol. 1, 138):

> In Quaternions, the square of a unit vector is -1. This singular convention is quaternionically convenient. But in the practical vector analysis of physics it

> is particularly inconvenient, being indeed, an obtrusive stumbling-block. All positive scalars products have the minus sign prefixed; there is thus a want of harmony with scalar investigations, and a difficulty in readily passing from Cartesians to vectors and conversely. My notation on the other hand, is expressly arranged to facilitate this mutual conversion.

When he wrote the last paragraph, Heaviside was already fully immersed in the volatile arguments between Tait—self-appointed champion of quaternions—and the vectorial targets of his wrath, specifically Gibbs. The polemical context of the last observation becomes clearly evident by the following introductory remarks (EMT, vol. 1, 136):

> 'Quaternion' was, I think, defined by an American schoolgirl to be 'an ancient religious ceremony'. This was, however, a complete mistake. The ancients—unlike Prof. Tait—knew not, and did not worship Quaternions. The quaternion and its laws were discovered by that extraordinary genius Sir W. Hamilton.

This entertaining example of Heaviside's distinct love of absurd humor may also serve to emphasize the extreme care with which such remarks must be read. The last sentence, despite its proximity to an acidic joke, may well have been meant in earnest. On an earlier occasion, before the quaternionic debate reached such high tones, he wrote (vol. 2, 557):

> Nevertheless, apart from practical application, and looking at it from the purely quaternionic point of view, I ought to also add that the invention of quaternions must be regarded as a most remarkable feat of human ingenuity. Vector analysis, without quaternions, could have been found by any mathematician by carefully examining the mechanics of the Cartesian mathematics; but to find out quaternions required a genius.

No sarcasm precedes this observation. It concludes a matter-of-fact exposition of vector algebra that contains a level-headed assessment of the relative merits and disadvantages of quaternions. When in similar mood, Heaviside did not dwell so heavily on the negative sign of the scalar product. In some cases, the negative sign could actually be given a sensible interpretation. For example, the quaternionic working of the differential operator $\nabla = (i\,d/dx + j\,d/dy + k\,d/dz)$ obtains a simple meaning in Heaviside's preference for expressing the physical significance of this operator with the words convergence, divergence, and curl. For if A is the quaternion $(0, A_x, A_y, A_z)$, then the quaternionic product

$$\nabla\mathsf{A} = -\left(\frac{dA_x}{dx} + \frac{dA_y}{dy} + \frac{dA_z}{dz}\right) + i\left(\frac{dA_z}{dy} - \frac{dA_y}{dz}\right) + j\left(\frac{dA_x}{dz} - \frac{dA_z}{dx}\right) + k\left(\frac{dA_y}{dx} - \frac{dA_x}{dy}\right) \tag{5}$$

$$= -S\nabla\mathsf{A} + V\nabla\mathsf{A} \quad \text{in the Hamilton–Tait notation,} \tag{6}$$

$$\equiv \nabla\cdot\mathbf{A} + \nabla\times\mathbf{A} \quad \text{in modern, Gibbs–Heaviside algebraic notation,} \tag{7}$$

$$= \mathrm{conv}(\mathbf{A}) + \mathrm{curl}(\mathbf{A}) \quad \text{in Heaviside's preferred terms.} \tag{8}$$

For Heaviside, who was responsible for many terminological innovations both in physics and mathematics, terminology and formalism were never ends in themselves; he consistently insisted that they must be shaped by the meaning they are intended to serve. He was well aware of the ∇ operator and its formal properties, but he usually preferred to address it by explicit reference to the meaning of its application to vector fields. The scalar part of the quaternion product above expresses the convergence, or flow of flux into an infinitesimal space rather than the divergence, or flow out of it, and hence he expressed it as conv(**A**). In this case, the $(-)$ sign represents no difficulty, and accordingly Heaviside made nothing of it (vol. 1, 271). On this occasion, he pointed out the potentially more disturbing double use of the Hamiltonian indices i, j, and k, to mean sometimes space orientations, and sometimes rotations about given axes.

In his early discussions of vector algebra, Heaviside tended to emphasize that quaternions were important because in them, for the first time ever, the three-dimensional space vector was given center stage, instead of being hidden under a mass of Cartesian manipulations. At the same time, he noted, the quaternionic product does not lend itself to easy physical interpretation, while the physically meaningful and widely applicable scalar and vector products must be isolated from it artificially. From these observations he concluded quite naturally that it would be useful to establish vectors and their basic operations on their own, independently of quaternions. In his earliest explicit reference to the need of creating an algebra of vectors (published in *The electrician* in December 1882) he expressed all of these concerns at length (vol. 1, 207):

> In mathematical investigations relating to electromagnetism, it often happens that the equations assume such a very complex form that the real meaning of the relations expressed by them becomes hidden away, as it were, beneath a tangled mass of x, y, z's, and can only be recognised by groping about from one equation to another [...] A very remarkable system of mathematics was invented by Sir W. Hamilton, called Quaternions, which may be described as the calculus of vectors. Owing to the universal presence of vectors in physical science, it is exactly fitted to express physical relations. Instead of breaking up vectors into three components, working with them as scalars, and then, when required, compounding them again to get back to vectors, (a most roundabout method), in the calculus of vectors we may fix our attention upon the vectors themselves, and work with them direct. [...] the calculus of Quaternions ought, then, one would say, to speedily supplant the ordinary methods in physical applications; in fact, it should have done so already. But it has not. Does this arise from mere Conservatism—the hatred of having to leave the old ways even for better? Although this may be partly true, it cannot be the whole truth. Against the above stated great advantages of Quaternions has to be set the fact that the operations met with are much more difficult than the corresponding ones in the ordinary system, so that the saving of labour is, in a great measure, imaginary. There is much more thinking to be done, for the mind has to do what in scalar algebra is done almost mechanically. At the same time, when working with vectors by the scalar system, there is great advantage to be found in continually bearing in mind the fundamental ideas of the vector system. Make

> a compromise: look behind the easily-managed but complex scalar equations, and see the single vector one behind them, expressing the real thing.

Tait was not the only one who objected to the new algebra, and even among those who considered that emphasis on vectors is useful, not all found Heaviside's treatment of them attractive. FitzGerald, one of very few individuals who succeeded to befriend the reclusive Heaviside, wrote to him in on 4 February 1889 (Heaviside Collection, Institute of Electrical Engineers, London): 'I am rather sorry you have not been content to work with the ordinary quaternions notation. It makes a very great difficulty to many people who want to look over and pick out the points in your work'. On 26 September 1892 he wrote again:

> I hope you will succeed in making the ordinary mathematical physicists think in vectors although I do not think your notation an improvement. You see I was very 'big' on Tait and get very much [bothered] by your omission of S [in front of a scalar product] and when one gets bothered every time one naturally takes a dislike to the botheration.

Despite such early difficulties, vector algebra became a sweeping success in short order. Nowadays, university programs in physics and mathematics routinely include a course on vector algebra. Quaternions, on the other hand, are much less well known, although there are some modern enthusiasts.

3 HEAVISIDE'S OPERATIONAL CALCULUS

Heaviside developed his operational calculus piecemeal for well over 20 years from the late 1880s. Just as his development of vector algebra reflects the desire to find a natural language for the discussion of force fields, so the operational calculus reflects a desire for treating the differential equations that arise in circuit and field theory in a way that reflects their physical meaning. A practical mathematician, he always put high value on the ability to obtain explicit solutions for the equations of mathematical physics. For this reason he always admired the power contained in the Fourier series approach to the solution of partial differential equations (§26). He felt, however, that all too often one loses the physical meaning of the equations amidst Fourierian manipulations (vol. 2, 389–390):

> Whilst it is impossible not to admire the capacity possessed by solutions in Fourier series to compactly sum up the effect of an infinite series of successive solutions, it is greatly to be regretted that the Fourier solutions themselves should be of such difficult interpretation. Perhaps there will be discovered some practical way of analysing them into easily interpretable form.

In the operational methods that he slowly developed, Heaviside found a partial remedy for such concerns. To properly appreciate this, a brief discussion of simple circuits is required.

The analysis of any electrical circuit is guided by two basic principles, usually referred to as Kirchhoff's circuit laws: 1) The sum of all currents entering a circuit element must be equal to the sum of all currents leaving the circuit element, or simply, the sum of all currents entering and leaving a circuit element must be zero; and 2) The sum of all voltage drops

over any closed loop must be zero. The first law is the electrical analogue of the principle of matter conservation, stating that in all interactions electrical charge is conserved; the second is analogous to Newton's third law of motion, with voltage drops playing the role of active forces. The dynamical analogy between electrical systems and mechanical ones is a central notion that guided much of Heaviside's work on the theory of electrical circuits and electromagnetic fields.

The basic elements of all circuits are voltage and/or current sources, resistors, inductors (coils), and capacitors. Voltage and current sources act as active agents that attempt to push current through a circuit; resistors, inductors, and capacitors are reactive elements, that impede the development of current and voltage, each in its own peculiar way. Generally, then, according to Kirchhoff's second law, when a voltage source is coupled to a circuit, opposed voltages will appear over the reactive elements of the circuit so that together they exactly counterbalance the impressed voltage. The formulation of this general statement in explicit electrical terms requires the relationship between current and voltage over the components that make up the circuit. Using Heaviside's designations, where the voltage and current in a circuit are V and C, while the resistance, inductance, and capacitance are R, L, and S, these relationships are

$$V_R = -R \cdot C, \quad V_L = -L \cdot dC/dt \quad \text{and} \quad C = -S\, dV_S/dt. \tag{9}$$

The negative signs are there to signify that the voltage drops over the elements are opposed to the direction of current pushed by the impressed voltage.

The simple circuit in Figure 1 contains an inductor (coil), a capacitor, and a resistor, all in series with a voltage source of intensity E. Kirchhoff's first law requires that the same current, C, flows through each of the elements at any given time. His second law requires that

$$E + V_L + V_S + V_R = 0, \tag{10}$$

where E is the impressed voltage (generally a function of time). Using the relationships for current and voltage over the individual elements, this may be turned into a differential equation for the voltage over the capacitor at any time, namely:

$$-E = V_S + RS\, dV_S/dt + LS\, d^2V_S/dt^2. \tag{11}$$

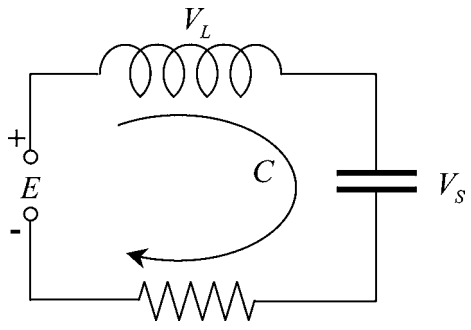


Figure 1. Circuit diagram.

This equation obtains a specific solution once the state of the capacitor is specified for a given point in time (for example, at $t = 0$, both current and voltage over the capacitor are also 0). Solutions to such equations were well known by Heaviside's time, and he employed them regularly whenever required.

Also normal was the practice of abbreviating the differential operator by a single letter, and treating it as 'algebraic', namely as if it were a mere number, with the usual commutative and associative properties (compare §36.2 on George Boole). Heaviside always followed this practice, treating it as no more than mere shorthand, an aid to memory in the sense of his observation (vol. 1, 196) that

> [...] as for the use of symbols, they are merely a sort of shorthand to assist the memory, which even those who openly contemn mathematical methods are glad to use so far as they can make them out—in the expression of Ohm's law for instance, to avoid spinning a long yarn.

Heaviside made this observation at the beginning of his first series of papers dedicated to the study of electromagnetic field dynamics. The mention of Ohm's law, however, is strangely prophetic of how a mere shorthand developed in Heaviside's hands into a calculus of operators. Using p to denote the time derivative, the defining relationships for resistance, inductance, and capacitance may be written as

$$V_R = -R \cdot C, \quad V_L = -Lp \cdot dC \quad \text{and} \quad C = -Sp \cdot V_S. \tag{12}$$

Treating p as 'algebraic', to use Heaviside's words, the capacitance relationship may be rewritten as $V_S = -(Sp)^{-1} \cdot C$; and now, Kirchhoff's voltage law for the simple circuit above becomes

$$E = (R + Lp + 1/Sp) = Z(p) \cdot C. \tag{13}$$

or alternatively,

$$C = E/Z(p), \tag{14}$$

$Z(p)$ is not a normal function, for it contains the operations of differentiation as well as the inverse of differentiation, the latter not yet properly defined. Ignoring this for a moment, the mere form of the operational voltage law appears like a generalized Ohm's law, relating voltage to current through a 'generalized resistance' for which Heaviside later coined the term 'impedance operator'. Ohm's law was the oldest and most familiar principle to the telegraphists and electricians among whom Heaviside began his career in electrical science and engineering. They knew that when resistors were connected in series, their combined resistance, R_t, could be calculated as $\sum_i R_i$, where the summation extends through all the individual resistors. Equally well known was that when connected in parallel, the total resistance could be obtained from

$$1/R_t = \sum_i 1/R_i. \tag{15}$$

Now consider two circuits, each containing a combination of resistors, capacitors, and inductors, with operational impedances Z_1 and Z_2 (both of which are functions of p, the

differential operator). The two circuits may be connected together with a voltage source either is series, or in parallel. From the basic Ohm's law like property of operational impedances, and by straightforward application of Kirchhoff's laws it follows that for the series connection $Z_t = Z_1 + Z_2$, and that for the parallel connection, $Z_t^{-1} = Z_1^{-1} + Z_2^{-1}$. It turns out, then, that operational impedances combine precisely like regular resistances. In the *Philosophical magazine* for December 1887, Heaviside published his first attempt to put together an introduction to what was gradually becoming for him a calculus of operators. He stressed the latter property of operational impedances (vol. 2, 355):

> The resistance-operator Z is a function of the electrical constants [...] and of d/dt, the operator of time-differentiation, which will in the following be denoted by p simply. As I have made extensive use of resistance-operators and connected quantities in papers, it will be sufficient here, as regards their origin and manipulation, to say that resistance-operators combine in the same way as if they represented mere resistances. It is this fact that makes them of so much importance, especially to practical men, by whom they will be much employed in the future. I do not refer to practical men in the very limited sense of anti- or extra-theoretical, but to theoretical men who desire to make theory practically workable by the simplification and systematisation of methods which the employment of resistance-operators and their derivatives allows, and the substitution of simple for more complex ideas.

Heaviside had reasonable cause for stressing the importance of the formal similarity between operational impedances and standard resistances. Proper use of this similarity could help to streamline writing down the characteristic equations of various circuits. However, practically minded people eventually desire explicit solutions of such equations, and as yet the operational formulation did not help at all. As he testified, he had been using resistance operators to express the equations of various circuits long before he wrote the exposition in 1887. To solve these equations he used Fourier series or Bessel functions depending on the specifics of the problem (for example, vol. 2, 176–177). Had this remained the state of affairs, one may doubt that the 1887 paper would have been written. Its publication is probably due to the sentence immediately following the above quotation regarding resistance operators and their derivatives: 'In this paper I propose to give a connected account of most of their important properties, including some new ones, especially in connection with energy'.

The new property referred to is what later became known as 'Heaviside's expansion theorem'. It transformed the formal operational expression of differential equations into a systematic means for extracting their explicit solutions. Motivation for the expansion theorem appears to reside again in the generalized Ohm's law that operational impedances associate with electrical circuits. In a circuit containing standard resistance only, if the impressed voltage is given, then the current at any time may be calculated by simple division. Now consider a voltage source E impressed at $t = 0$ upon a circuit containing capacitive and inductive elements in addition to standard resistive ones. If a specific mathematical meaning could be associated with the formal division of E by the operational impedance $Z(p)$, then the operational Ohm's law of this circuit would become the explicit solution

for the current at any subsequent time. In the expansion theorem Heaviside found what appeared to be the beginning of a general answer to this question.

Demonstration of the expansion theorem is too lengthy for this brief account; proofs may be found in texts on the operational calculus. The remarkable thing about Heaviside's original discussion, is that the expansion theorem emerges out of energy considerations pertaining to the analysis of electrical circuits (vol. 2, 372–373). The form of the expansion theorem that he presented in his 1887 paper states that if a steady voltage source E is impressed at $t = 0$ on an electrical circuit characterized by operational impedance $Z(p)$, then the current at any subsequent time is given by

$$C(t) = \frac{E}{Z(0)} + \sum_{i=1}^{N} \frac{E}{p_i \frac{dZ}{dp}|_{p=p_i}} e^{p_i t}, \tag{16}$$

where the p_i are the algebraic roots of $Z(p) = 0$. Here p is treated as a complex number, not as the operation of deriving with respect to time, and the expression yields C as an explicit function of time. In (16) the expansion theorem is limited to cases where $Z(p)$ can be written as $\prod_i (p - p_i)$. If higher powers of any of the expressions in parentheses exist, then the expansion theorem is not valid in the form (16). Heaviside showed later how to generalize the expansion theorem to include such cases as well.

In principle, then, the expansion theorem provides a mechanical procedure for obtaining the solution of the circuit's differential equation. An additional advantage is that the expansion solution automatically takes account of the initial conditions that must be dealt with separately in the usual manner of solving the differential equation. In practice, however, application of the expansion theorem proves tedious, and only common sense and experience can help to decide whether its use for specific problems is advantageous.

Heaviside may have appreciated these limitations, for when he wrote 'On electromagnetic waves' (art. 43 above), he adopted a different approach. He wrote the field equations operationally, expressed the solutions in symbolic operational form, and then re-expressed them as combinations of Bessel functions with operational arguments. Great care needed to be taken in so doing, because the operational arguments could not be taken to behave like mere numbers under all circumstances (vol. 2, 446). Heaviside showed that certain physical conclusions can be extracted from these expressions while still in their operational form. Then he gave some examples of how the series may be converted, element by element, into explicit functions of time and space. Key to this procedure was the realization that when the operator $\sqrt{p}$ operates on a function whose value is 1 for $t \geqslant 0$ and 0 otherwise, the result is $(\pi t)^{-1/2}$ (see, for example, the note to vol. 2, 446–447).

4 ON HEAVISIDE'S LATER WORK

After *Electrical papers* was completed, Heaviside devoted much time in the 1890s to develop further his operational calculus. The main product of this endeavor was a series of three long papers 'On operators in physical mathematics' that were intended for publication by the Royal Society (to which Heaviside had been elected Fellow in 1891). Only the first two were published: the third was turned down in the face of stern objections

raised by mathematicians to his unconventional and potentially problematic use of divergent series. He eventually published the full mathematical content of all three papers in the second volume of his compendium *Electromagnetic theory* (dedicated to the memory of FitzGerald), amidst many caustic remarks about closed-minded Cambridge mathematicians [Heaviside, 1899]. From the formal point of view, his most important addition to the work that he began in the *Electrical papers* was the use of the unit function (now generally referred to as 'the Heaviside unit function' or 'unit') to resolve the non-commutative nature of differentiation and integration, effectively turning the latter into a proper algebraic inverse of the differential operator p [Yavetz, 1995, appendix 4.2]. Overall, however, his quest for a complete operational calculus proved far more difficult than the formulation of vector algebra, and in some ways remained elusive [Lützen, 1979].

When all is said and done, Heaviside did not succeed in tying all the elements of his operational calculus into a tight formal framework as he did with vector algebra. He did, however, succeed to provide through it an innovative and powerful set of techniques for solving differential and partial differential equations with constant coefficients. His work on the operational calculus motivated a great deal of subsequent research in this field of practical and applied mathematics.

BIBLIOGRAPHY

Aristotle 1953. *Minor works* (trans. W.S. Hett), Cambridge, MA: Harvard University Press.

Crowe, M.J. 1967. *A history of vector analysis: the evolution of the idea of a vectorial system*, Notre Dame: University of Notre Dame Press. [Repr. New York: Dover, 1985.]

Heaviside, O. 1893, 1899, 1912. *Electromagnetic theory*, 3 vols., London: The Electrician Publishing Company. [Repr. in 1 vol. London: Spon, 1951. Also repr. New York: Chelsea, 1971.]

Lodge, O. 1884. 'On the seat of the electromotive forces in the voltaic cell', *The telegraph journal and electrical review*, *15*, 365–368, 380–382, 407–410. Also in *Report of the fifty-fourth Meeting of the British Association for the Advancement of Science, 1884* (1885), 464–529.

Lützen, J. 1979. 'Heaviside's operational calculus and the attempts to rigorize it', *Archive for history of the exact sciences*, *21*, 161–200.

Maxwell, J.C. 1892. *A treatise on electricity and magnetism*, 2 vols., 3rd ed. Oxford: Clarendon Press. [Repr. New York: Dover, 1954. See §44.]

Nahin, P.J. 1988. *Oliver Heaviside: sage in solitude*, New York: IEEE Press.

Tait, P.G. 1867. *Elementary treatise on quaternions*, Oxford: At the University Press. [2nd ed. 1873.]

Yavetz, I. 1995. *From obscurity to enigma: the work of Oliver Heaviside, 1872–1889*, Basel: Birkhäuser.

CHAPTER 50

WALTER WILLIAM ROUSE BALL, *MATHEMATICAL RECREATIONS AND PROBLEMS OF PAST AND PRESENT TIMES*, FIRST EDITION (1892)

David Singmaster

In recent years, it has become recognized that recreational mathematics is an interesting branch of mathematics with a long and fascinating history, revealing much about mathematics and popular culture. Ball's book was one of the first substantial books devoted to recreational mathematics.

First publication. London: Macmillan, 1892. 12 + 240 pages.

Later editions. 2nd–11th are London: Macmillan. 2nd ed., May 1892 (with several changes in chs. XI and XII). 3rd ed. retitled *Mathematical recreations and essays*, 1896, 12 + 276 pages. 4th ed. 1905, 16 + 388 pages. 5th ed. 1911, 16 + 492 pages. 6th ed. 1914 ('Chapter XVIII has been re-written'). 7th ed. 1917 (Ch. XI replaced by two new chapters). 8th ed. 1919. 9th ed. 1920. 10th ed. 1922, 14 + 366 pages; repr. 1926, 1928, 1931, 1937. 11th ed. (rev. H.S.M. Coxeter), 1939, 16 + 418 pages; repr. 1940, 1942, 1944, 1947, 1956, 1959, 1962, 1963, 1967. 12th ed. (rev. Coxeter), Toronto: University of Toronto Press, 1974, 18 + 428 pages. 13th ed. (rev. Coxeter: very few changes), 18 + 428 pp., New York: Dover, 1987.

French translation. *Récréations et problèmes mathématiques des temps anciens et modernes* (trans. J. Fitz-Patrick from the 3rd ed. 1896, 'Revue et augmentée par l'auteur'), 1st ed., Paris: Hermann, 1898. 2nd ed. (from the 4th ed., additions by A. Hermann, Fitz-Patrick, and others), 3 vols., 1907–1909. [Repr. with changes 1926–1927. Photorepr. in 1 vol. Paris: Gabay, 1992. Vols. 1 and 2 available on-line at http://gallica.bnf.fr.]

Italian translation. *Ricreazioni e problemi matematici dei tempi antichi e moderni* (trans. Dionisio Gamboli from the 4th ed. 1905, with some additions in ch. IV), Bologna: Zanichelli, 1910.

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

German translation? A letter from 'An old pupil' after Ball's death says the book was translated into German; but I have found no trace of this, and [Ahrens, 1910–1918] does not mention it.

1 HISTORICAL BACKGROUND

Recreational problems are scattered through the mathematical literature from the beginnings of recorded mathematics. The first general work to include a number of mathematical puzzles is *The Greek anthology*, compiled by Metrodorus about 510, which includes 44 simple problems. There are no answers in the manuscripts. The problems include 'aha' problems (for example, Pythagoras's age), cistern problems, and ass and mule problems [Singmaster, 1984–1985].

The first collection of mathematical recreations is the *Propositiones ad acuendos juvenes*, a manuscript reasonably attributed to Alcuin of York around 800, though it was also attributed to Bede. Alcuin [c.800] has 53 numbered problems with answers (the Bede version has three extra ones). Some problems have several answers, but only one is given and there are no real explanations for any of the problems. Several problems occur here for the first time ever: e.g. the river crossing problems of the man with a wolf, a goat and some cabbage; that of the three jealous couples; or, for the first time in Europe, the hundred fowls problem. It is clear that the author has compiled his problems from earlier sources which we generally do not know.

Several major mathematical works have devoted much space to problems that are now considered recreational, notably the following, which are available in various versions:

- *Chiu Chang Suan Ching* (*Nine chapters on the mathematical art*) (around 150).
- Aryabhata (I), *Āryabhaṭīya* (499);
- Mahavira, *Gaṇita-sāra-sangraha* (850);
- Bhaskara (II), *Bijaganita* and *Lilavati* (both 1150);
- Leonardo Fibonacci, *Liber abbaci* (1202, though all extant copies are from the second edition of 1228); and
- Luca Pacioli, *Summa de arithmetica geometria proportioni & proportionalita* (1494).

Traditionally, books on mercantile arithmetic included a number of miscellaneous recreational problems, including simple divinations.

The first large work devoted to recreational mathematics is another work of Luca Pacioli, his *De viribus quantitatis* [Pacioli, c.1500], which is only now receiving the attention it deserves. This is a manuscript of 618 pages in Bologna, apparently compiled during 1496–1509. Agostini [1924] described the 81 arithmetic problems in Part 1. Part 2 contains 134 geometrical problems, including some topological puzzles. Part 3 contains several hundred proverbs, poems, riddles and tricks (that is, physical recreations, conjuring, etc.). There is no standard English version of the title: I suggest *On the powers of numbers*. Though the manuscript is clearly written, the microfilm version is sometimes faint.

The text presents the usual difficulties of manuscripts of the time. Many of the geometric problems refer to diagrams, which are missing. I find it very difficult to understand the geometric problems, and only around 1998 did I realise that Pacioli gives the earliest known examples of several classical topological puzzles. Much of the difficulty has been rectified by the publication of a transcription of the text, but this omits the unique marginal drawing of a string puzzle on folio 206r and the transcription has some errors. Dario Uri has photographed the entire manuscript and enhanced the images to produce a more legible version of the text on a CD. He has discovered that these problems include the earliest known discussion of the Chinese Rings, previously first known to be in Cardano, as well as about a dozen other such puzzles in their earliest known forms. Bill Kalush has discovered several earliest examples of magic tricks. We now feel this is definitely the earliest recreational mathematics book—except that it was never published.

Girolamo Cardano and Niccolo Tartaglia include much recreational material in their works in the mid 16th century. The first known book on conjuring appeared in 1584 and contained some mathematical divinations and some topological puzzles. From this time a flood of conjuring books has appeared, and many contain some recreational mathematics.

The first recreational mathematics book to be published was [Bachet, 1612], which also starts the tradition of longevity of such books—it is still in print! It has only 35 arithmetical problems and a few others like the river crossing problems, but it deals with them in some detail and with some mathematical technique and notation.

Following and copying much from Bachet and also from Tartaglia and Cardano is [van Etten and Leurechon, 1624]. The authorship is considerably disputed: van Etten's name is on the book, but since 1643 it has been attributed to Jean Leurechon, who may have been van Etten's teacher. It might have been a joint effort. This book started another tradition, passing through many versions—at least 67 in four languages between 1624 and 1706—and has been attributed to about eight different authors as well as 'anonymous'. My bibliography of it occupies 19 pages, so I cannot include it here.

The book is an example of another tradition in the field: it includes many physical recreations. For example, it is thought to be the first to think of a telegraph, based on the specious belief that two magnetic needles would always stay aligned, and the first use of the word thermometer. It also includes optical and magnetic phenomena, geographical and astronomical puzzles, an ear trumpet, perpetual lamps, tricks with sound, games, aeolipiles (a pneumatic instrument), colossal statues, giants, dialling, ballistics, etc., etc. As can be seen, the material is very miscellaneous, and there is little attempt to explain any mathematics.

From this point on, there are many works which cover mathematical and physical recreations. The best known are those of Schwenter, Schott, Witgeest, Ozanam, Guyot and Hooper. But these tended to be miscellaneous collections with little coherence except that some group their problems in sections, such as Arithmetic, Geometry, Optics, Dialling, Cosmography, Mechanicks, Physicks, Pyrotechny. Of these only Ozanam [1694] attempts to systematise the material and to use mathematical notation and thinking. His books also exemplifies the longevity and multiplicity of recreational books: there were at least 29 editions in two languages (my bibliography on this occupies 10 pages), and an English version was reissued as late as 1854. In about 1723, the work was expanded to four volumes and much interesting material was added, particularly a section on conjuring and puzzles, but

this material was deleted in a new four-volume version in 1778. Ozanam also is the most extreme example of another common property of recreational books: it was massively pirated and copied.

In the mid 19th century, a massive number of books for boys (and sometimes girls) and for general amusement appeared in England, typical titles being *The boy's own book*; *The girl's own book*; *The boy's own conjuring book*; *The magician's own book*; *Philosophical recreations, or, winter amusements*; *Endless amusement*; and *Rational amusement for winter evenings*. They all include sections on mathematical puzzles and games, generally fairly randomly arranged, with minimal solutions or mathematics. The impetus for this explosion of recreational books may have come from France, since one of the books states it is a translation from *Le magicien des salons*, but it is not clear which work that is. There are several similar French works, but the examples that I have seen contain very little in the way of mathematical puzzles and recreations.

2 THE LATE 19TH CENTURY

It is not until the end of the 19th century that fresh approaches occur, in France, Germany and England, with the works of Lucas [1882–1894], Ball's own book from 1892, [Hoffmann, 1893] and [Schubert, 1898]. The work of Ahrens [1910–1918] rounds off this period. Let us survey these works before examining Ball.

2.1 (*François-*) *Édouard* (*-Anatole*) *Lucas* (*1842–1891*) was a teacher at various *lycées* and produced only a few papers until 1875, when he began to put out about a dozen papers per year on number theory and geometry. In 1879 he began publishing articles on games and started a popular series of 13 articles in *Revue scientifique* (1879–1883), followed by 16 articles in *La nature* (1886–1890). They formed the basis of his four-volume work [Lucas, 1882–1894]. He died unexpectedly in his prime: at a banquet, a piece of a dropped plate scratched his cheek and he died five days later of blood poisoning. He inspired a generation of recreational mathematicians in France, much in the way that Martin Gardner was to do in the 20th century. He also had historical interests and was an editor of Pierre Fermat's works.

In Lucas's book each chapter is based on a single idea. He develops many topics not previously covered in any book, and thoroughly systematizes and generalizes them. Volume 1 (1882) has only eight chapters, whose titles I paraphrase: River crossing problems; 'The game of bridges and islands' (basic graph theory, starting with the Bridges of Königsberg); Labyrinths; The eight queens problem, etc.; Solitaire; Binary numeration; Chinese rings; and Taquin (the fifteen puzzle). His other volumes include chapters on Kirkman's schoolgirls, Hamilton's icosian game, Calculating machines, Mathematical games, Roulette, Perpetual calendars, The four colour problem, and Walking machines. Much of the material is relatively recent and he frequently adds new data, regularly citing colleagues who have made suggestions and improvements. But he was also a bit of a prankster, and he may have invented some of these colleagues. For example, he seems to have invented the story that the Chinese rings were used to lock chests in Norway, attributed some of the ideas to 'ex-students'; and he never admitted his invention of the Tower of Hanoi in print.

At the end of volume 1, Lucas has an 'Index bibliographique' of 11 pages listing 190 items, starting with one 15th-century item. However, he never mentions publishers, hardly

ever gives first names and often omits the author's first initial. His knowledge is sometimes a bit vague—he cites [van Etten and Leurechon, 1624] seven times under five different authors and twice anonymously! Each volume has a detailed table of contents, but no index.

2.2 'Professor Louis Hoffmann' was the pseudonym of Angelo John Lewis (1839–1919), a London barrister with an interest in conjuring. He wrote a series of articles on magic for *Every boy's annual* that were collected as *Modern magic* (1876), which is considered the foundation of modern conjuring. He went on to write about two dozen books on magic, games, puzzles and recreations, becoming the best known author on these subjects in late-19th-century England.

Hoffmann [1893] is the first substantial book devoted to puzzles. The chapters are, in my paraphrasings: Dexterity puzzles; Trick or secret puzzles; Dissections; Arithmetical puzzles; Word puzzles; Puzzles with counters; Matchstick puzzles; Wire puzzles; 'Catch' puzzles; Miscellaneous puzzles. The mechanical trick and wire puzzles are real novelties, many of them having first appeared in the late 19th century. There are only a few references, and these are to suppliers of puzzles. There is a detailed table of contents, but no index. Hoffmann gives a short description of 'Elementary properties of numbers', which even asserts that any prime of the form $(4k + 1)$ is a sum of two squares, and he uses straightforward algebra to solve the arithmetic problems; but the book really marks the division of mechanical puzzles from mathematical recreations.

2.3 Hermann Cäsar Hannibal Schubert (*1848–1911*) was a secondary school teacher in Hamburg. His work on enumerative geometry was the subject of one of Hilbert's problems of 1900, which was given rigorous proofs in 1912 and 1930 (§57). He also edited a series of textbooks for school and lower university use, and compiled collections of problems. In 1891–1894 he produced a series of columns on recreational topics in the *Naturwissenschaftlichen Wochenschrift*. He collected them in *Zwölf Geduldspiele* (1895) and then expanded it into *Mathematische Mussestunden* [Schubert, 1898]. It has gone through at least 13 editions, the last(?) being in 1967. The second edition appeared in three volumes in 1900 and in an abbreviated one-volume form in 1904. Both versions had third editions, in 1907–1909 and 1907 respectively. Later editions were one-volume works, but steadily expanded in size. Ahrens says he has seen another printing of the three-volume form [1918, vol. 2, 417].

Mathematische Mussestunden is a good workman-like book, with considerable coverage. Schubert rarely gives any references, his bibliography contains only 14 items, and there is no index and only a brief table of contents; all of this lessens its value to the historian. I find it a bit dry, but its long popularity in Germany shows it was suitable for many readers.

2.4 Wilhelm Ernst Martin Georg Ahrens (*1872–1927*) spent a few years teaching at higher schools, but in 1904 he retired to his native Rostock and devoted himself to writing. The first edition of his *Mathematische Unterhaltungen und Spiele* appeared in one volume in 1901. The second edition comprised two volumes, the first published in 1910, but the second delayed until 1918.

Ahrens's book is the most detailed history of recreational mathematics. He is knowledgeable about early printed books, medieval and earlier writings and has some knowledge of Arabic and Japanese material, and he gives detailed references. For example, his discussion of the history of the Josephus problem occupies 30 pages and he then devotes 23 pages to its mathematics. His bibliography is 57 pages with 762 items, starting with [Alcuin, c.800]. He gives details of later editions, but not publishers. He even lists the 22 items that he had not personally seen and wondered if they existed—and at least six do. Ahrens had the immense advantages of following Lucas, Schubert and Ball and of not having to teach, so he was able to go far beyond his predecessors. He has a very comprehensive 16-page index of names, including all the authors in the bibliography, and an eight-page index of topics. In view of the immense value of this book, it is amazing that it has never been reprinted. In 1988, the publisher Teubner told me that they were planning to do so, but it has not yet happened.

3 WALTER WILLIAM ROUSE BALL (1850–1925)

Now we turn to our author. Walter William Rouse Ball was born in Hampstead, London on 14 August 1850, as the only son of Walter Frederick Ball. It is not known when or why he adopted the style 'Rouse Ball'. He attended University College School and University College London, where he won the gold medal in Mathematics and first-class honours in Logic and Moral Philosophy in 1869. He entered Trinity College Cambridge in 1870, and was Second Wrangler and First Smith's Prizeman in 1874. He then went to study law at the Inner Temple in London and was called to the Bar there, but only practised as an equity draftsman and conveyancer. However he wrote *The student's guide to the Bar* (1878), which went through at least seven editions. He was elected a Fellow of Trinity in 1875. He deputised for W.K. Clifford at University College London in 1877.

Rouse Ball returned to Trinity as Lecturer in 1878 and remained until (semi-)retiring in 1905. He served as Assistant Tutor, Director of Mathematical Studies, Tutor, Senior Tutor, Chairman of the College Education Committee and Secretary of the College Council. Around 1919 he founded the Pentacle Club for conjuring. *Granta* once advertised a new game called 'Rous-ball'. Ball was devoted to making Trinity a great centre of mathematics: after he became Senior Tutor in 1898, he organized the administration to give free rein to researchers. Bertrand Russell, G.H. Hardy, James Jeans, E.T. Whittaker, Arthur Eddington, J.E. Littlewood, G.W. Watson, G.I. Taylor, G.H. Darwin and Sydney Chapman were some of the products of his era.

Ball married Alice Mary in 1885, who died in December 1919; apparently there were no children. Whittaker [1925] says that Ball and his wife soon won the reputation of being the best tutor and tutor's wife around. The Balls built the house, 'Elmside', 20 (now 49) Grange Road, in about 1890. Still called 'Elmside', it is now used by Clare Hall as a residence. Ball built a maze in his garden, which is described and illustrated in *Mathematical recreations and essays*, from the fourth edition (1905) onwards; but there is now no trace of the maze. Ball's editor H.S.M. Coxeter told me that the maze was only made of posts and strings, which were changed every few weeks so each student had a different maze to run. Ball died in the house on 4 April 1925, and was buried in the Ascension Burial Ground (formerly St. Giles' Cemetery) in Huntingdon Road.

Ball began writing about the time that he came to Cambridge, starting with the aforementioned student's guide in 1878. Almost all of his 15 some books deal with the history of mathematics or Cambridge or Trinity College; but he also wrote an *Elementary algebra* (1890), and *An Introduction to string figures* (1920), based upon a lecture to the Pentacle Club. His *A short account of the history of mathematics* (1880) went through at least six editions and was translated into French, Italian and Spanish. With J.A. Venn he also produced five volumes of *Admissions to Trinity College, Cambridge* (1911–1916).

Ball made an extensive collection of portraits of mathematicians, perhaps the largest in the world. It was exhibited at the Napier Tercentenary Exhibition in Edinburgh, and a catalogue appears in the *Handbook of the Napier tercentenary celebration* (1914). The albums of these photographs are in Trinity Library, catalogued as 'Adv. Alb. 2–10'. They look like a most interesting resource and it would be splendid to have them published. A wall-plaque commemorates him in Trinity College Antechapel, but few of his papers are in the College Library.

4 THE PUBLICATION OF BALL'S BOOK

The contents by editions of the book are displayed in Table 1. The bibliography is more complicated than initially appears. [van Etten and Leurechon, 1624] and [Ozanam, 1694] nominally have more editions; but most of these are reprintings of the same text, and there are really only about six forms of the first work and three of the second. By contrast, Ball lived until after the 10th edition and he made changes in almost every one. Even between the fifth and the ninth editions, where the basic structure of the book remained fairly static and the number of pages did not change, there were usually some substantial changes of content. One exception is the ninth edition, which appears to be a reprint of the eighth with only one deletion of about half a page. So the historian needs to examine each edition up through the 12th—to find, for example, that the river crossing material has six different forms! The French translation is considerably augmented, both by Ball and by others, so it also needs to be examined.

Though a competent historian, Ball only gives footnote references in this book. These are quite explicit and thorough, but one wishes for a bibliography. The table of contents is reasonably detailed and there is a moderate index. Like Lucas, he suffered from the fact that non-European mathematics was not well known in Europe, despite the work of many orientalists who had already translated much of the basic material. Ball generally cites only back to [Bachet, 1612], sometimes mentioning Tartaglia; and he cites [van Etten and Leurechon, 1624] only once. Under the three Greek classical impossible problems and under mazes, he cites various Greek and Latin authors, usually from late editions. Only in the material on π does he make any references to Arabic, Indian and Chinese material, or even to Fibonacci. Overall, Ball is not as useful as Ahrens for tracing older history, but his book is invaluable for its period, as we will now see.

5 EXAMPLES OF NEW MATERIAL IN THE BOOK

5.1 General Kayles. One has a row or a circle of objects and one can remove one object or two (or perhaps more) adjacent ones. Two players alternate and the one who removes the

Table 1. Contents by editions of Ball's book.
The book divides into two parts, at 'The Mathematical Tripos'. The numbers indicate the number of pages devoted to the chapter (or chapters) devoted to the topic. Some material has been shuffled around between chapters so that this table does not always reflect the pages devoted to a topic. Each unlisted edition is essentially the same as its predecessor.

Editions	**1**	**3**	**4**	**5**	**7**	**8**	**10**	**11**	**12**
Arithmetical recreations.	28	37	39	42	42	42	42	75	74
Geometrical recreations.	24	34	36	39	39	39	39	53	54
Polyhedra.								32	32
Mechanical recreations.	19	22	26	26	26	26	26		
Chess-board recreations.				28	28	28	28	32	31
Magic squares.	14	19	26	32	32	24	24	29	29
Bees and their cells.						8	8		
Map-colouring problems.								20	21
Unicursal problems.	28	30	31	23	23	23	23	26	28
Kirkman's school-girls.				31	31	31	31	32	
Combinatorial designs.									41
Miscellaneous problems.	34	34	38	23	23	23	40	27	26
The Mathematical Tripos.			39	36					
Calculating prodigies.					28	28	28	29	28
Calculating machines.					8	8			
Three classical problems.	23	23	24	23	23	23	23	24	22
The parallel postulate.				20	20	20			
Insolubility of the quintic.				6	6	6			
Mersenne's numbers.			15	15	15	15			
String figures.				32	32	32	16		
Astrology.	16	16	16	15	15	15			
Cryptographs and ciphers.			30	29	29	29	29	32	31
Hyper-space.	12	13	13	14	14	14			
Time and its measurement.	17	20	21	21	21	21			
Matter and aether theories.	15	18	22	23	23	23			
Index.	7	8	10	10	10	10	7	8	10
Total.	240	276	388	492	492	492	366	418	428

last object wins. It is based on an old physical game where one throws a stick at a row of pins. As a mathematical game, it seems to originate about 1897 with the puzzlers S. Loyd or H.E. Dudeney, but Ball's fifth edition (1911) seems to be the first to propose the general version with p counters in a circle and one can take up to m adjacent counters.

5.2 Exploration problems. These are problems of getting into or across a desert where one can carry more food than one needs for a day, so one can pass on food to another, or one can leave caches of food at depots to be picked up later. We have recently observed that versions of this appear in [Alcuin, c.800] and [Pacioli, c.1500]. Again, Ball's fifth edition (1911) is the first with a systematic study. He distinguishes two forms of the problem, with n explorers who can carry food for d days:

a) Without depots, they can get one man $nd/(n+1)$ days into the desert and back.

b) and more common: with depots permitted, they can get a man $d(1/1+1/2+\cdots+1/n)/2$ days into the desert and back.

5.3 Fore and aft puzzle. Consider the part of a 5×5 board consisting of two 3×3 subarrays at diagonally opposite corners. They overlap in the central square. One square has 8 black men and the other has 8 white men, with the centre left vacant. One can move a man horizontally or vertically toward the opposite corner, either by moving to an adjacent empty space or by jumping over a man of the opposite colour to an empty space. This is given in Ball's first (1892) and third (1896) editions, where he states that he believes he was the first to publish the puzzle, but 'that it has been since widely distributed in connexion with an advertisement and probably now is well known'. He drops this assertion in the fifth edition (1911). Hoffmann [1893] says the puzzle is on sale in the United Kingdom and that there is an 1894 US patent, but I have not found any mention of it before 1892.

5.4 Chessboard placing problems. In the third, fourth and fifth editions Ball initiates a number of questions and extends previous questions on the maximum or minimum number of pieces one can place on a chessboard under various restrictions; that is, generalizations of the eight queens problem. These problems are still being actively studied.

5.5 Geometric fallacies. In his first edition Ball says he seems to be the first to publish 'Every triangle is isosceles' and 'A right angle is obtuse'.

5.6 Variant of the Josephus problem. The original problem is to arrange n good guys and n bad guys in a circle and count off by k so that the bad guys are eliminated. Dudeney gave examples where different counts and starting points would eliminate either the good guys or the bad guys. In his fifth edition (1911) Ball asks if this can happen with different counts and the same starting point and gives examples for $n=2,3,4$; for example, G B G G B G B B, counted by 5s and 9s. The 10th edition (1920) gives a solution for any n due to a Mr. Swinden.

5.7 Salary puzzle. It is better to get a rise of £5 every half year than £20 every year. This appears in the 3rd edition (1896) as a question 'which I have often propounded in past years', and I myself have found no earlier mention of it.

5.8 Magic tours. Ball's fifth edition (1911) seems to be the first to give a magic square of order 8 where the numbers form a king's tour.

5.9 Water in wine versus wine in water. If one takes a spoonful of water and adds it to a container of wine, then takes a spoonful of the mixture and returns it to the water container, is there now more water in the wine or wine in the water? In the third edition (1896) Ball says this is a question 'which I have often propounded in past years'. I have no earlier specific reference, but a reminiscence of Lewis Carroll (1832–1898) by Viscount Simon says it was a favourite problem with him. Simon entered Wadham College Oxford, in 1892 and met Carroll after that time.

5.10 1089. Take a three-digit number, with first digit smaller than its last, and subtract it from its reversal. Then add the result to its reversal and you get 1089. The first statements of this fact in this decimal version seem to be in Ball's French edition (1898) and in the fourth edition (1905). Surprisingly, the English monetary version, solved by £12 18s 11d, preceded this—for example in the first edition, where Ball cites an 1890 appearance which says it has been 'current in well-informed City-circles for some months' and gives a general solution for any monetary system of three levels. I have now seen another 1890 version of the £ s d version. (I believe that I have seen a reference around 1881, but I cannot trace it.) The French translator of Ball adds that the result in base b is $(b-1)(b+1)^2$. Carroll's nephew, S. Dodgson Collingwood, writing in 1899, thought Carroll invented the problem in the £ s d form, which is possible.

6 CONCLUDING REMARK

Like Lucas, Schubert and Ahrens, Ball was greatly involved with creating new problems and in reporting on new problems from colleagues. He was also fortunate in living for a long time. These points account for the constant changes in the book and its wide coverage, which make it especially valuable for the history of this period. These facts and Ball's fluent and friendly presentation are the keys to the book's immense popularity. It has inspired several generations of English-speaking school students to take up mathematics and, in Coxeter's sympathetic revisions, it remains as popular now as it was when it first appeared.

BIBLIOGRAPHY

Agostini, A. 1924. 'Il "De viribus quantitatis" di Luca Pacioli', *Periodico di matematiche, (4) 4*, 165–192.

Ahrens, W.E.M.G. 1910–1918. *Mathematische Unterhaltungen und Spiele*, 2nd ed. 2 vols., Leipzig: Teubner.

Alcuin c.800. *Propositiones ad acuendos iuvenes.* [Can be found in the works of Alcuin and Bede, for example in the *Patrologiae Latinae*. English trans. by John Hadley and David Singmaster as 'Problems to sharpen the young', *Mathematical gazette*, *76* (1992), 102–126; corrected and updated edition available from me.]

Bachet, C.-G.B. 1612. *Problèmes plaisans & délectables qui se font par les nombres*, Lyon: P. Rigaud. [2nd ed. 1624.]

Cajori, F. 1926. 'Walter William Rouse Ball', *Isis*, *8*, 321–324.

Hoffmann, L. [pseudonym of Angelo John Lewis]. 1893. *Puzzles old and new*, London: Warne. [Repr. with Foreword by L.E. Hordern, London: Martin Breese, 1988. Corrected, with colour photos, by Hordern, as *Hoffmann's puzzles old & new*, Cane End: Hordern, 1993.]

Lucas, (F.-) É.(-A.) 1882–1894. *Récréations mathématiques*, 4 vols., Paris: Gauthier–Villars. [2nd eds. of vol. 1 (1891) and vol. 2 (1893). Repr. Paris: Blanchard several times.]

Metrodorus. c.510. In *The Greek anthology* (trans. W.R. Paton), London: Heinemann, 1916–1918 (Loeb Classical Library), vol. 5, Book 14.

Ozanam, J. 1694. *Recreations mathematiques et physiques, qui contiennent Plusieurs Problémes* [sic] *utiles & agreables, d'arithmetique, de geometrie, d'optique, de gnomonique, de cosmographie, de mecanique, de pyrotechnie, & de physique. Avec un traité nouveau des horloges elementaires*, 2 vols., Paris: Jombert. [Various later eds.]

Pacioli, L. c.1500. *De viribus quantitatis*, Italian manuscript in Biblioteca Universitaria, Bologna, Codex 250. [Probably compiled 1496–1509. Microfilm available. Transcription by Maria Garlaschi Peirani, preface and ed. Augusto Marinoni, Milan: Ente Raccolta Vinciana, 1997.]

Schubert, H.C.H. 1898. *Mathematische Mussestunden*, Leipzig: Göschen.

Singmaster, D. 1984–1985. 'Puzzles from the Greek Anthology', *Mathematical spectrum, 1*, no. 1, 11–15.

Singmaster, D. 1998. 'The history of some of Alcuin's *Propositiones*', in P.L. Butzer and others (eds.), *Charlemagne and his heritage. 1200 years of civilization and science in Europe*, vol. 2, *Mathematical arts* (ed. H.Th. Jongen and W. Oberschelp), Turnhout: Brepols, 11–29.

van Etten, H. and Leurechon, J. 1624. *Recreation mathematicque. Composee de plusieurs problemes plaisants et facetieux. En faict d'arithmeticque, geometrie, mechanicque, opticque, & autres parties de ces belles sciences*, Pont-a-Mousson: Jean Appier Hanzelet. [2nd ed. 1626.]

Whittaker, E.T. 1925. 'W.W. Rouse Ball', *Mathematical gazette, 12*, 449–454.

CHAPTER 51

ALEXANDR MIKHAILOVICH LYAPUNOV, THESIS ON THE STABILITY OF MOTION (1892)

J. Mawhin

This memoir is recognized as the first extensive treatise on the stability theory of solutions of ordinary differential equations. It is the source of the so-called Lyapunov first and second methods.

First publication. *Ob'shchaya zadacha ob'ustoichivosti dvizheniya* [*The general problem of stability of motion*], Kharkov: Kharkov Mathematical Society, 1892. 250 pages. [Doctoral dissertation, University of Kharkov.]

Second edition. Moscow and Leningrad: Academy of Science, 1935. [With a portrait, additions from the 1907 French version, Russian translation of [Lyapunov, 1897] and an obituary by V.A. Steklov.]

Third edition. Moscow and Leningrad: GITTL, 1950. [With a portrait, and the papers [Lyapunov, 1893a, 1893b, 1897].]

Fourth edition. As *Collected works*, vol. 2, Moscow: Academy of Science, 1956, 7–263. [Contains all published papers of Lyapunov on the stability of solutions of ordinary differential equations, and an unpublished list of the seven theses adjoined to the dissertation.]

All the above editions appeared in Russian.

French translation by E. Davaux, 'Problème général de la stabilité du mouvement', *Annales de la Faculté des Sciences de Toulouse, (2) 9* (1907), 203–474. [Revised and corrected by the author, with additional note. Repr. Princeton: Princeton University Press, 1949 (Annals of Mathematics Studies, no. 17); also Paris: J. Gabay, 1988.]

English translation. The general problem of the stability of motion (trans. A.T. Fuller), in *International journal of control, 55* (1992), no. 3 (Lyapunov centenary issue). [Also published separately, London: Taylor and Francis, 1992. Contains editorial by Fuller, the biography [Smirnov, 1992] and bibliography [Barrett, 1992].]

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

Related articles: Lagrange on mechanics (§16), Thomson and Tait (§40), Poincaré (§48), Birkhoff (§68), Volterra (§73).

1 THE AUTHOR

Alexandr Mikhailovich Lyapunov was born in 1857, the son of the astronomer Mikhail Vasilievich Lyapunov, who worked at Kazan University before becoming the director of a Lyceum in Yaroslavl. Lyapunov's brother Sergei was a composer; another brother, Boris, was a specialist in Slavic philology and became a member of the Soviet Academy of Science.

Lyapunov received his elementary education at home before graduating from the Gymnasium of Nizhny Novgorod and entering at the Physics and Mathematics Faculty of Saint Petersburg University, where P.L. Chebychev greatly influenced him. He graduated in 1880 and obtained his master's thesis in 1884 on *The stability of ellipsoidal forms of equilibrium of a rotating liquid*. He taught mechanics as a *Privatdocent* at Kharkov University and published there in 1892 his classical memoir *The general problem of the stability of motion* (in Russian), defending it the same year as a doctoral dissertation at Moscow University.

In 1893 Lyapunov became a professor at Kharkov and made researches on mathematical physics, in particular on the Dirichlet problem, and the calculus of probability. In 1901, he was elected as a member of the St. Petersburg Academy of Science, taking the seat that had remained vacant for seven years since the death of Chebychev. In 1917, with the hope of improving the health of his wife, who suffered from a serious form of tuberculosis, Lyapunov moved to Odessa, where he taught at the university. But his wife died on 31 October 1918, and he shot himself, surviving his wife by only three days. For more biographical information, see [Grigorian, 1974; Smirnov, 1992].

Lyapunov's work on the stability of solutions of ordinary differential equations started with his doctoral dissertation of 1892 (subsequently referred as *Dissertation*) and covered a period of ten years. The nine other contributions, listed in [Barrett, 1992], give a few additions to the general theory of stability and substantial complements to the study of linear second order equation with periodic coefficients.

2 THE AIM AND THE INSPIRATION OF THE *DISSERTATION*

The object of Lyapunov's *Dissertation* is clearly indicated in the *Preface*:

> In this work are exposed some methods for the resolution of questions concerning the properties of motion and, in particular, of the equilibrium, which are known under the denominations of stability and instability [...]. The problem consists in knowing if it is possible to choose the initial values of the solutions x_s small enough so that, for all values of time following the initial instant, those functions remain, in absolute value, smaller than limits given in advance, as small as we want. When we can integrate our differential equations, this problem does not present real difficulties. But it would be important to have methods which would allow to solve it, independently of the possibility of this integration [...].

Then he analyzes and criticizes the 'linearization method' usually adopted in stability questions, since the pioneering work of J.L. Lagrange, P.S. Laplace (§18.4) and S.D. Poisson, by authors like W. Thomson and P.G. Tait (§40), E.J. Routh, and N.E. Zhukovski:

> The procedure usually used consists in neglecting, in the considered differential equations, all the terms of order greater than one with respect to the quantities x_s and to consider, instead of the given equations, the linear equations so obtained. [...] But the legitimacy of such a simplification is not justified a priori and [...] if the solution of the simplified problem can give an answer to the original one, it is only under certain conditions, which, generally, are not indicated.

Then Lyapunov mentions his principal source of inspiration:

> The unique tentative, as far as I know, of rigorous solution of the question belongs to M. Poincaré, who, in a remarkable memoir *'Sur les courbes définies par les équations différentielles'*, and in particular in the last two parts, considers stability questions for differential equations of the second order as well as close questions relative to systems of the third order. Although M. Poincaré restricts himself to very special cases, the methods that he uses allow much more general applications and can still provide many new results. This is what will be shown in what follows because, in a large part of my researches, I have been guided by the ideas developed in the quoted Memoir.

Finally, Lyapunov explicits the aim of his *Dissertation*:

> The problem that I have posed to myself, in starting the present study, can be formulated as follows: to indicate cases where the first approximation really solves the stability question, and to give procedures which would allow to solve it, at least in some cases, when the first approximation is no more sufficient.

3 LYAPUNOV'S CONCEPT OF STABILITY

The contents of Lyapunov's *Dissertation* are summarised in Table 1. The first Chapter, entitled 'Preliminary analysis', contains precise definitions of the used concepts and the development of the general methods applied in the two subsequent chapters. The solution with initial value x_0 at initial time t_0 of the ordinary differential system (written here, in contrast to Lyapunov, in vector notation)

$$dx/dt = X(x, t), \tag{1}$$

is denoted by $x(t, t_0, x_0)$.

To define and study the concept of *stability* of a solution $\xi(t)$ of (1), Lyapunov first observes that the substitution $x \to \xi + x$ reduces the question to the stability of the zero solution of a system of the type (1) satisfying $X(0, t) \equiv 0$. He calls this zero solution *stable* if for each $\varepsilon > 0$ and each t_0, one can find $\eta > 0$ such that for each x_0 with $\|x_0\| \leqslant \eta$ and

Table 1. Contents by Sections of Lyapunov's dissertation.
The numbers of pages in the rows for the Chapters refer to the French/English translations (1907 and 1992) respectively; in other rows Section numbers are given. DE = differential equations.

Sections	**Topics and methods**
6/4	Preface. Concepts of stability and instability. Earlier work: Thomson–Tait, Routh, Joukovsky, Poincaré. Summary of the memoir.
58/51	*Chapter I. Preliminary analysis*.
1–5	Generalities on the considered question. Stability, unstability. Solutions of DE by power series.
6–10	On some systems of linear DE. Characteristic numbers. Normal systems. Regular systems.
11–13	On a general case of DE of perturbed motion. Convergent series solutions of DE. The first method.
14–16	Some general propositions. Positive and negative definite functions. The second method.
124/110	*Chapter II. Study of steady motions.*
17–21	Linear DE with constant coefficients. Construction of a Lyapunov function. Canonical systems.
22–41	DE of the perturbed motion. Sufficient conditions for stability and instability. Inversion of the Lagrange–Dirichlet stability theorem. Linearization with one zero root or two imaginary roots.
42–45	Periodic solutions of the perturbed motion.
72/72	*Chapter III. Study of periodic motions*.
46–47	Linear DE with periodic coefficients. Floquet theory.
48–53	Some propositions on the characteristic equation. Second-order equation. Canonical systems.
54–64	Study of the DE of the perturbed motion. Sufficient conditions for stability and instability. Linearization with one characteristic factor equal to one. Linearization with two imaginary characteristic factors of modulus one.
65	A generalization.

all $t \geqslant t_0$, one has $\|x(t, t_0, x_0)\| < \varepsilon$. This is essentially the continuous dependence of the solution on initial conditions, *for all values of* t larger than the initial one.

The zero solution is called *unstable* if it is not stable. The related concept of *uniform stability*, in which η is independent of t_0, was to be introduced in 1933 by K.P. Persidskii. At the end of Chapter I, Lyapunov gives a refinement of the concept of stability, called today *asymptotic stability*, in which, in addition to stability, he requires that $x(t, t_0, x_0) \to 0$ when $t \to +\infty$ for each t_0 and each sufficiently small $\|x_0\|$.

After having proved, by the method of majorants, the existence of convergent series for the solutions of (1) of sufficiently small norm, defined over an arbitrary interval of time, Lyapunov introduces a term still used today, although maybe in a slightly more restricted sense: the set of all procedures of study of the stability depending upon the rendition of solutions of the perturbed motion in the form of infinite series, is called the *first method*.

The *second method* consists in all types of procedures which are independent of obtaining solutions of the differential equations of the perturbed motion.

4 THE FIRST METHOD OF LYAPUNOV

As one can write $X(x,t) = P(t)x + R(x,t)$, where $R(x,t) = \mathrm{O}(\|x\|^2)$, the linear system

$$dx/dt = P(t)x \tag{2}$$

is called the *linearization* or the *variational equation* of (1) around the zero solution. The first step consists in studying the stability of its trivial solution, in order to deduce possible information on the stability of the trivial solution of (1).

For this, Lyapunov introduces the concept of *characteristic number* of a function $x(t)$ such that $x(t)\exp\lambda_1 t \to 0$ and $x(t)\exp\lambda_2 t \to \infty$ as $t \to +\infty$, for some λ_1 and λ_2. Then, a number λ_0 exists such that, for each $\varepsilon > 0$, $x(t)\exp(\lambda_0 + \varepsilon)t \to \infty$ and $x(t)\exp(\lambda_0 - \varepsilon)t \to 0$ when $t \to +\infty$. λ_0 is called the *characteristic number* of the function $x(t)$. An equivalent definition

$$\lambda_0 := \lambda(x, \exp t) = -\lim_{t\to+\infty} \sup \big(\log|x(t)|/t\big), \tag{3}$$

has been given in 1930 by O. Perron, who proved that the set of characteristic numbers of the linear system (2) contains at most n distinct elements. The negative of the Lyapunov characteristic numbers and their analogues for discrete dynamical systems play, under the name of Lyapunov *exponents*, an important role in the recent researches on chaos.

When P is constant or periodic, the sum of its characteristic numbers is equal to

$$-\lim_{t\to+\infty} \sup (1/t)\int_{t_0}^{t} \Re\big[\operatorname{tr} P(\tau)\big]\,d\tau \tag{4}$$

and Lyapunov calls *regular* a system satisfying this condition. Their study has been continued by O. Perron, N.G. Cetaev and Persidskii. An important subclass of regular systems introduced by Lyapunov are the *reducible systems*, that is, systems (2) which can be reduced to a system with constant coefficients through a transformation of the type $x = Q(t)y$, where $Q(t)$ is of class C^1, bounded on $[t_0, +\infty[$ together with the determinant of its reciprocal. They have been studied by N.P. Erugin and I.Z. Shtokalo.

Lyapunov is then ready to state and prove the basic theorem of his first method: *If the linearized system is regular and if all its characteristic numbers are positive, then the unperturbed motion is stable, and moreover the perturbed motion tends asymptotically to the unperturbed one when t tends to $+\infty$.*

5 THE SECOND METHOD OF LYAPUNOV

Lyapunov then proceeds to his second method, whose aim is to extend the Lagrange–Dirichlet stability theorem to not necessarily conservative systems. In his words (Sec. 16):

> Everybody knows the theorem of Lagrange on the stability of equilibrium in the case where a potential exists, as well as the elegant proof which has been proposed by Lejeune-Dirichlet. This last proof rests upon considerations which can be used to prove many other analogous theorems.

J.P.G. Lejeune-Dirichlet (1805–1859) had proved in 1846, by qualitative arguments, that an equilibrium of an autonomous conservative mechanical system is stable if it is a strict minimum of the potential function V. Lagrange's earlier proof, based upon linearization, was insufficient (compare §16).

After introducing and analyzing the concept of positive definite or negative definite function $V(x,t)$ (a *positive definite* $V(x,t)$ is bounded below by a continuous increasing function $\varphi(\|x\|)$ vanishing at 0), Lyapunov proves his fundamental result: *the trivial solution of system* (1) *is stable if one can find a definite function* $V(x,t)$ *whose derivative along solutions of* (1)

$$\left\langle V'_x(x,t)\middle| X(x,t)\right\rangle + V'_t(x,t) \tag{5}$$

has a fixed sign opposite to that of V, *or is identically zero, in some neighborhood of the origin.* The idea of the very simple proof goes back to Dirichlet, and consists, given $\varepsilon > 0$ and t_0, in taking $\eta > 0$ such that $V(t_0, x_0) < \varphi(\varepsilon)$ whenever $\|x_0\| < \eta$. As $V(t, x(t, t_0, x_0))$ is nonincreasing, assuming the existence of a first $t_1 > t_0$ such that $\|x(t_1, t_0, x_0)\| = \varepsilon$ leads to a contradiction.

Lyapunov also observes that if, in addition, V *has an infinitesimal upper bound and a defined derivative along solutions of* (1), then the zero solution is asymptotically stable. He also proves in this setting two sufficient conditions for *instability*, in terms of properties of some Lyapunov functions. They will be refined by many authors, starting with Cetaev in 1934.

Those types of functions V are nowadays called Lyapunov *functions*, and a lot of energy has been used to find ways of constructing them. Much emphasis has been put also on finding suitable types of stability implying the existence of a suitable Lyapunov function (*converse theorems*), in the hands of Persidskii, I.G. Malkin, J.L. Massera, J. Kurzweil, N.N. Krasovskii and E.A. Barbashin. The second method of Lyapunov, which is also useful to study various types of *asymptotic behavior of solutions* of differential equations, is often referred as Lyapunov's *direct method*.

6 THE CASE OF AUTONOMOUS SYSTEMS

In Chapter two ('Study of steady motions'), Lyapunov applies his second method to the special case where the linear approximation has constant coefficients. He first reproves the simple case where the stability or instability follows from the linear approximation. Incidentally, he rediscovers independently some results of Poincaré's Ph.D. thesis of 1879.

Lyapunov observes (Sec. 35) that the Lagrange–Dirichlet theorem on the stability of a mechanical systems in the presence of a potential

> gives a sufficient condition for stability, consisting in the fact that the potential must reach a minimum at the equilibrium position. But, in proving that this

> condition is sufficient, this theorem does not allow to conclude to the necessity of the same condition. This is why the following question can be raised: will the equilibrium position be unstable if the potential is not minimum? In this general form, this question is not solved up to now. But, under some assumptions of rather general character, one can answer it in a precise way.

After a century of research, despite substantial advances, the situation of this problem can still be described exactly in Lyapunov's words, except when V is analytical, for which case V.M. Palamodov has proved in 1995 the converse of the Lagrange–Dirichlet theorem. See [Rouche et alii, 1977] and [Hagedorn and Mawhin, 1992] for references.

Lyapunov then analyzes in detail the situations where the characteristic equation of the linear approximation has one zero root and the other ones have negative real parts, and the case where it has two purely imaginary roots, the other ones having negative real parts. Those cases are known nowadays as *critical*, for the linear approximation is no more sufficient to decide of the stability of the trivial solution. He has considered the case of two zero roots for the characteristic equation in a manuscript which has only been published posthumously [Lyapunov, 1963], and completed by V.A. Pliss in 1964. One finds in Lyapunov's treatment the germ of the theory of *center manifolds*, a fundamental tool for many contemporary researches on ordinary differential equations and dynamical systems.

On this occasion, Lyapunov also states and proves his famous theorem on the *existence of a family of periodic solutions near the origin in the presence of a first integral*. Consider an autonomous differential system

$$dx/dt = X(x), \tag{6}$$

where X is analytic, $X(0) = 0$, $X(0)$ has a pair of imaginary eigenvalues $\alpha_1 = i\omega$, $\alpha_2 = -i\omega$ for some $\omega > 0$ and the other eigenvalues such that α_k/α_1 is not an integer for $3 \leqslant k \leqslant n$ (*nonresonance condition*). Assume moreover that the system (6) admits a first integral G with non-vanishing Hessian on the space E spanned by the eigenfunctions associated to $\pm i\omega$. Then Lyapunov proves that *for each sufficiently small ε there exists a unique T-periodic solution $x(t;\varepsilon)$ near E with period $T(\varepsilon)$ close to $2\pi/\omega$ lying in the set $G(x) - G(0) = \varepsilon^2$ and such that $x(t;\varepsilon) \to 0$ and $T(\varepsilon) \to 2\pi/\omega$ as $\varepsilon \to 0$.*

An example of J. Moser shows that the non-resonance condition on the α_k is necessary, but in 1973 A. Weinstein proved that it is superfluous in the case of a Hamiltonian system with the Hessian of the Hamiltonian definite at zero. *Global* versions of the Lyapunov theorem on families of periodic solutions have been obtained in the 1980s for the Hamiltonian case, following Rabinowitz in 1982, who used modern techniques of critical point theory. References on the modern local and global developments of Lyapunov's work on periodic solutions can be found in [Starzhinskii, 1977; Mawhin and Willem, 1989]. Those results are important in celestial mechanics.

7 THE CASE OF PERIODIC SYSTEMS

The last chapter of Lyapunov's monograph ('Study of periodic motions') concentrates on the case where the system (1) depends periodically on t. Then its linearized system (2) has

periodic coefficients, say of period ω. He starts by recalling the classical Floquet theory for such systems, stating (except for the matrix notations) that the principal matrix solution of (2) (for which $Y(0) = I$) can always been written in the form

$$Y(t) = Q(t)e^{tM}, \tag{7}$$

for some ω-periodic nonsingular matrix $Q(t)$ and some constant matrix M, whose characteristic roots are called the *characteristic exponents* of (2). Consequently, for all values of t,

$$Y(t+\omega) = CY(t), \quad \text{where } C = e^{\omega M}, \tag{8}$$

and the characteristic roots of C are called the *characteristic multipliers* of (2). Their explicit determination is of course in general impossible, but Lyapunov proves a number of their properties, in particular that *the characteristic multipliers of the adjoint system to* (2) *are the reciprocal of the characteristic multipliers of* (2). He also finds useful information on the characteristic multipliers when the coefficients of the system satisfy some symmetry conditions, and when (2) is Hamiltonian.

Lyapunov also initiates the study of the *second-order linear equation*

$$y'' + p(t)y = 0, \tag{9}$$

with ω-periodic coefficient $p(t)$, and finds explicit conditions upon p providing important information on its characteristic multipliers. He proves that *if* $0 \neq p \leqslant 0$, *the characteristic multipliers of* (9) *are real, one larger than one and the other one smaller than one*. On the other hand, if $0 \neq p \geqslant 0$, and if

$$\omega \int_0^\omega p(t)\,dt \leqslant 4, \tag{10}$$

the characteristic multipliers of (9) *are imaginary and have modulus one*. Those results have been generalized and refined by many authors, including O. Haupt, G. Hamel, G. Borg, I.M. Gelfand, V.B. Lidskii, M.G. Krein, V.A. Yakubovich, V.M. Starzhinskii and H. Hochstadt. Many refinements of Lyapunov *inequality* (10) have been obtained [Yakubovich and Starzhinskii, 1972].

Finally, Lyapunov combines his general results of Chapter 1 with his studies of linear periodic systems to prove that, *when* (1) *is ω-periodic in t, its trivial solution is asymptotically stable when all the characteristic multipliers of its linearization have moduli strictly smaller than one, and is unstable if one of them has modulus strictly larger than one*. Like in the autonomous case, he also discusses in length some difficult *critical* cases, where one characteristic multiplier is equal to one or where two characteristic multipliers are imaginary and of modulus one.

8 THE INFLUENCE OF POINCARÉ'S WORK ON LYAPUNOV'S *DISSERTATION*

We have seen that, in the preface of his *Dissertation*, Lyapunov generously acknowledges Poincaré's influence. In a footnote to his preface, he quotes Poincaré's King Oscar Prize

memoir *Sur le problème des trois corps et les équations de la Dynamique* (1890) (§48), as well as the first volume of the *Méthodes nouvelles de la Mécanique céleste* (1892), just published during the printing of the *Dissertation*. Describing later in the preface his method of development of solutions of ordinary differential in power series, Lyapunov mentions in a footnote that

> the series under study have been considered, under special conditions, in my memoir 'Sur les mouvements hélicoïdaux permanents d'un corps solide dans un liquide' (Communications de la Société mathématique de Kharkow, 2e série, t. I, 1888). I have learned after that M. Poincaré had considered those series, under the same hypotheses, in his Thesis 'Sur les propriétés des fonctions définies par les équations aux différences partielles' (1879).

This connection is made explicit in Chapter 2, Sec. 24. In this chapter, Lyapunov underlines the pioneering contributions of Poincaré in his series of memoirs 'Sur les courbes définies par une équation différentielle' (1881–1886), to what is called to-day the problem of determining the conditions under which an equilibrium of a planar differential system is a *center*, i.e. is surrounded by a one-parameter family of closed orbits. Other results of this series of memoirs are also mentioned in Sec. 64 of Chapter 3. Furthermore, in a footnote ending Sec. 45 of Chapter 2, devoted to periodic solutions, Lyapunov observes:

> The question of periodic solutions of nonlinear differential equations is also considered, although with another viewpoint, in the last memoir of Poincaré: 'Sur le problème des trois corps et les équations de la Dynamique' (Acta Mathematica, t. XIII).

In Chapter 3, when he states his theorem that a linear canonical system with periodic coefficients has a reciprocal characteristic equation, Lyapunov mentions in a footnote of Sec. 51 that

> his theorem is also indicated by M. Poincaré in his memoir 'Sur le problème des trois corps et les équations de la Dynamique' (Acta Mathematica, t. XIII, p. 99–100) [. . .]. But I knew it before the publication of this memoir and, in February 1900, I have communicated it, in the previous form, at the Mathematical Society of Kharkow, with other propositions related to the characteristic equation (Communications de la Société mathématique de Kharkow, 2e série, t. II; report of the meetings).

Summarizing, we see that Lyapunov's work has been influenced by Poincaré's, and often overlaps with Poincaré's further contributions. On several occasions, and specially in dealing with the question of stability, Lyapunov transforms into powerful general methods some remarks, made by Poincaré in special situations.

If there is common material in the work of Poincaré and Lyapunov, many differences attend their approach and style. Poincaré's insight is mostly geometrical, and Lyapunov's one essentially analytical. Further, it is striking to see how organized is Lyapunov's *Dissertation*, in contrast to [Poincaré, 1892–1899], which is a patchwork of descriptions of tools and results, with an amazing and wide scope. The comparison between those two giants

of differential equations is somewhat reminiscent of that between Bernhard Riemann and Karl Weierstrass in their approaches to complex function theory (§34); their followers have taken advantages of both styles.

9 THE EARLY RECEPTION OF THE WORK OF LYAPUNOV ON STABILITY

The third volume of Emile Picard's famous *Traité d'analyse* [Picard, 1893–1896] is almost entirely devoted to the study of differential equations. In Chapter VIII he describes Poincaré's theory of periodic solutions, with a somewhat more extended discussion of the existence of periodic solutions of an autonomous differential system *around an equilibrium*. Some of his incorrect conclusions were mentioned to Picard by Lyapunov, in a letter of 20 January 1895 (reproduced as Appendix III of [Mawhin, 1994]) in which Lyapunov provides a nice summary in French of his *Dissertation*, and informs Picard about his own results on periodic solutions. Lyapunov notices that Picard's reasoning about the existence of periodic solutions near a situation of equilibrium is not conclusive, except in the presence of a first integral. In 1897, Picard presents to the French *Académie des Sciences* a note of Paul Painlevé, which exhibits a counter-example to Picard's claim, but again proposes too optimistic an existence condition. Lyapunov is not mentioned. Despite Lyapunov's and Painlevé's remarks, the sections devoted to the periodic solutions near an equilibrium remain unaltered in the subsequent editions of Picard's *Traité*.

In the second edition (1908) of Volume III of his *Traité*, however, Picard adds a section to Chapter VIII entitled 'De la stabilité et de l'instabilité des intégrales de certaines équations différentielles; théorème de M. Liapounoff sur l'instabilité de l'équilibre'. He refers only to a note of Lyapunov published in Liouville's journal [Lyapunov, 1897], summarizing some of the concepts and results of the *Dissertation*, and giving new instability conditions based upon the second method. In his work 'Sur certaines propriétés des trajectoires en Dynamique', crowned by the *prix Bordin* of the *Académie* in 1896 and published the next year in the same issue of Liouville's journal as Lyapunov's note, Jacques Hadamard (1865–1963) studies the stability and asymptotic behavior of the trajectories of a mechanical system, through auxiliary functions similar to Lyapunov's ones. Hadamard mentions that the condition he has found for the instability of the equilibrium of a conservative mechanical system, was obtained by Lyapunov in 1892, in a memoir 'unfortunately written in Russian, an extract of which having been published in the journal of Jordan in 1897, whose existence was unknown to me when I communicated the above remarks to the Académie des Sciences'. See [Mawhin, 1994] for more details and references.

In Italy, T. Levi-Civita already mentions Lyapunov's memoir in a paper of 1897 which criticizes, like Lyapunov, the work of the British school based upon unjustified linearization. Inspired by classical mechanics, Levi-Civita requires, in the Lyapunov-like definition of stability introduced in his main work of 1901, that the conclusion holds for *all* values of t, and not only in the future. He calls this concept the *unconditional Dirichlet-type stability*, to distinguish it from Lyapunov's. Another important aspect of Levi-Civita's contributions is his detailed study of the stability of 'transformations', i.e. of mappings, anticipating the modern theory of dynamical systems [Dell'Aglio and Israel, 1989].

10 THE LATER DEVELOPMENT OF LYAPUNOV STABILITY

The contributions of Lyapunov to stability were considered important enough by French mathematicians to be included in some of their traditional large treatises on analysis published in Paris by Gauthier–Villars, such as the second edition (1910–1915) of Edouard Goursat's famous *Cours d'analyse mathématique*. (The last treatise in this tradition seems to be the *Cours d'analyse de l'Ecole Polytechnique* of J. Favard, 1960–1963.) After Bourbaki's influence, Lyapunov stability theory was expelled from general treatises of analysis; but it is found, besides the specialized monographs, in most books on ordinary differential equations.

In the former Soviet Union, the interest in Lyapunov theory seems to start around 1930, with the work of Cetaev on instability, of Persidskii on the first method and of Malkin on the second method. The first treatise on Lyapunov stability was published by Cetaev immediately after the Second World War [Cetaev, 1946], and has seen four editions. It has been followed, besides many research papers, by some sixty monographs on stability and its application to mechanics and control theory, among which one must mention the classics [Malkin, 1952; Letov, 1955; Zubov, 1957; Krasovskii, 1959; Aizerman and Gantmacher, 1963; Barbashin, 1967].

In the United States G.D. Birkhoff, Poincaré's brilliant follower, makes significant contributions to dynamical systems between 1912 and 1945 (§68), but Lyapunov's work on stability in only briefly mentioned in one or two memoirs. The introduction of Lyapunov theory there is due to a topologist and algebraic geometer of Russian origin, Solomon Lefschetz, who starts, during the Second World War, a new career devoted to differential equations and control theory. He creates a strong interest in stability theory among American mathematicians, as exemplified by the publication of about twenty monographs; following the first one, by Richard Bellman [Bellman, 1953], one should notice [Cesari, 1959; LaSalle and Lefschetz, 1961; Lefschetz, 1965; and Bhatia and Szegö, 1970]. In Europe and Japan, the books [Hahn, 1959; Yoshizawa, 1966; Rouche et alii, 1977] have been very influential and are now classical.

The techniques of Lyapunov have been successfully applied to other classes of equations, like integral or functional–differential equations, differential or evolution equations in Banach spaces, non-linear parabolic equations, and to discrete dynamical systems and difference equations. Lyapunov's techniques and results have important applications in mechanics, control theory, chaos theory, mathematical biology, population dynamics and economics. More than a century after its publication, Lyapunov's *Dissertation* remains an invaluable source of inspiration for mathematicians specialized in differential equations, dynamical systems and their applications. Its first English translation was published in 1992.

BIBLIOGRAPHY

Aizerman, M.A. and Gantmacher, F.R. 1963. *Absolute stability of control systems*, Moscow: Akademii Nauk. [In Russian. English trans. San Francisco: Holden-Day, 1964.]

Barbashin, E.A. 1967. *Introduction to the theory of stability*, Moscow: Nauka. [In Russian. English trans. Groningen: Wolters–Noordhoff, 1970.]

Barrett, J.F. 1992. 'Bibliography of A.M. Lyapunov's work', *International journal of control, 55*, 785–790.

Bhatia, N.P. and Szegö, G.P. 1970. *Stability theory of dynamical systems*, Berlin: Springer. [Repr. Berlin: Springer, 2002.]

Bellman, R. 1953. *Stability theory of differential equations*, New York: McGraw–Hill.

Cesari, L. 1959. *Asymptotic behavior and stability problems in ordinary differential equations*, Berlin: Springer. [2nd ed. 1963, 3rd ed. 1971.]

Cetaev, N.G. 1946. *Stability of motion*, Moscow: GITTL. [In Russian. 2nd ed. 1955, 3rd ed. 1965, 4th ed. 1990; Moscow: Nauka. English trans. Oxford: Pergamon, 1961.]

Dell'Aglio, L. and Israel, G. 1989. 'La théorie de la stabilité et l'analyse qualitative des équations différentielles ordinaires dans les mathématiques italiennes: le point de vue de Tullio Levi-Civita', *Cahiers du séminaire d'histoire des mathématiques, 10*, 283–321. [Published Paris: Université Pierre et Marie Curie.]

Grigorian, A.T. 1974. 'Lyapunov, Aleksandr Mikhailovich', in *Dictionary of scientific biography*, vol. 8, 559–563.

Hagedorn, P. and Mawhin, J. 1992. 'A simple variational approach to a converse of the Lagrange–Dirichlet theorem', *Archive for rational mechanics and analysis, 120*, 327–335.

Hahn, W. 1959. *Theorie und Anwendung der direkten Methode von Lyapunov*, Berlin: Springer. [English trans. Englewood Cliffs: Prentice Hall, 1963.]

Krasovskii, N.N. 1959. *Some problems of the theory of stability of motion*, Moscow: Fizmatgiz. [In Russian. English trans. Stanford: Stanford Univ. Press, 1963.]

LaSalle, J. and Lefschetz, S. 1961. *Stability by Lyapunov's direct method with applications*, New York: Academic Press.

Lefschetz, S. 1965. *Stability of nonlinear control systems*, New York: Academic Press.

Letov, A.M. 1955. *Stability of nonlinear control systems*, Moscow: GITTL. [In Russian. English trans. Princeton: Princeton University Press, 1960.]

Lyapunov, A.M. 1893a. 'On the problem of stability of motion', *Zapiski Imperatorskogo Khar'kovskogo Universiteta, 1*, 99–104; also *Soobshcheniya Khar'kovskogo Matematicheskogo Obshchestva, (2) 3*, 265–272. [In Russian.]

Lyapunov, A.M. 1893b. 'Investigation of one of the special cases of the stability of motion' , *Matematicheskii sbornik, 17*, 253–333. [In Russian.]

Lyapunov, A.M. 1897. 'Sur l'instabilité de l'équilibre dans certains cas où la fonction de forces n'est pas un maximum', *Journal de mathématiques pures et appliquées, (5) 3*, 81–94.

Lyapunov, A.M. 1963. *Investigation of one of the singular cases of the problem of stability of motion* (ed. V.P. Bassov), Leningrad: Leningrad University. [In Russian. English trans.: *Stability of motion* (with a contribution by V.A. Pliss), New York: Academic Press, 1966.]

Malkin, I.G. 1952. *Theory of the stability of motion*, Moscow: GITTL. [In Russian. 2nd ed. Moscow: Nauka, 1966. German trans. Berlin: Akademie Verlag; Munich: Oldenburg, 1959.]

Mawhin, J. 1994. 'The centennial legacy of Poincaré and Lyapunov in ordinary differential equations', *Rendiconti del Circolo Matematico di Palermo*, ser. II, no. 34, suppl., 9–46.

Mawhin, J. and Willem, M. 1989. *Critical point theory and Hamiltonian systems*, New York: Springer.

Picard, E. 1893–1896. *Traité d'analyse*, 3 vols., Paris: Gauthier–Villars. [2nd ed., 1901–1908.]

Poincaré, H. 1892–1899. *Les méthodes nouvelles de la mécanique céleste*, 3 vols., Paris: Gauthier–Villars. [Repr. Paris: Blanchard, 1987.]

Rouche, N., Habets, P. and Laloy, M. 1977. *Stability theory by Lyapunov's direct method*, New York: Springer.

Smirnov, V.I. 1992. 'Biography of A.M. Lyapunov', *International journal of control, 55*, 775–784.

Starzhinskii, V.M. 1977. *Applied methods in the theory of nonlinear oscillations*, Moscow: Nauka. [In Russian. English trans. Moscow: Mir, 1980.]

Yakubovich, V.A. and Starzhinskii, V.M. 1972. *Linear differential equations with periodic coefficients*, Moscow: Nauka. [In Russian. English trans. New York: Wiley, 1975.]

Yoshizawa, T. 1966. *Stability by Lyapunov's second method*, Tokyo: Mathematical Society of Japan.

Zubov, V.I. 1957. *The method of A.M. Lyapunov and their applications*, Leningrad: Leningrad University. [In Russian. 2nd ed. Moscow: Vyssh. Shkola, 1984. English trans. Groningen: Noordhoff, 1964.]

CHAPTER 52

HEINRICH HERTZ, POSTHUMOUS BOOK ON MECHANICS (1894)

Jesper Lützen

Offering a mechanical foundation of physics that avoided force as a basic concept, this was the first book on mechanics to make use of Riemannian geometry in configuration space. In the philosophical introduction Hertz described physical theories as (mental) images of the natural world.

First publication. *Die Prinzipien der Mechanik in neuem Zusammenhange dargestellt* (ed. P. Lenard), Leipzig: Barth, 1894. xxxii + 312 pages.

Manuscripts. Several drafts and the almost finished manuscript in Hertz's hand are preserved at the *Deutsches Museum*, München.

Reprint. As *Gesammelte Werke,* vol. 3, Leipzig: Barth, 1910.

Photoreprint. Vaduz: Sändig, 1984.

English translation. *The principles of mechanics presented in a new form* (trans. D.E. Jones and J.J. Walleye), London: Macmillan, 1899. [Photorepr. New York: Dover, 1956.]

Related articles: Newton (§5), d'Alembert (§11), Lagrange on mechanics (§16), Riemann on geometry (§39), Thomson and Tait (§40), Maxwell (§44), Hilbert on geometry (§55).

1 EDUCATION AND EMPLOYMENTS

Heinrich Hertz (1857–1894) was one of the last physicists who made lasting contributions to both theoretical and experimental physics (for a book-length biography see [Fölsing, 1997]). His unusual talents in both theoretical and practical matters were manifest already during his childhood. Top student in his school class, he was particularly gifted in mathematics and languages, especially Arabic, and in the afternoons he enjoyed to work at his carpentry bench and at his lathe. In an attempt to combine his theoretical and practical interests he matriculated at the Dresden Polytechnic in 1876 as a student of constructional

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

engineering. After one semester his studies were interrupted by a one-year draft to the railroad troops. When he resumed his studies at the Polytechnic in Munich he soon decided to switch to physics, and after one year there he moved on to the leading physics laboratory of the time, that of Hermann von Helmholtz (1821–1894) at the University of Berlin.

At this time Helmholtz was engaged in electromagnetic research. His main objective was to decide if electromagnetic phenomena could best be described by Wilhelm Weber's theory of action at a distance or by Clerk Maxwell's field theory (§44). Hertz soon embarked on experimental research that came out in favour of Maxwell's theory. It earned him a prize from the University. Simultaneously he followed courses in theoretical physics, including three in mechanics given by C.W. Borchardt, G.R. Kirchhoff, and Ernst Kummer. Having earned his doctorate with a thesis on the currents induced on a rotating sphere by a magnet, he continued to work as Helmholtz's assistant. While at Helmholtz's laboratory, he did experimental and theoretical research on a variety of subjects such as electromagnetism, elastic deformations, evaporation, the tides, a new dynamometer, floating plastic plates and cathode rays. This research resulted in 11 publications.

During the years 1883–1885 Hertz held a post as *Privatdozent* at the university in Kiel. Since there were no laboratory facilities at this small university, his research during this period on electromagnetism and hydrodynamics was purely theoretical. In the summer semester of 1884 he held a public series of lectures entitled 'Modern ideas on the constitution of matter'. His carefully written notes from these lectures have recently been published by Albrecht Fölsing [Hertz, 1999]. They give a very well informed and thoughtful survey of the contemporary ideas about the constitution of the ether and of ponderable matter; and they reveal that already at this time Hertz thought of theoretical explanations of microscopic physical phenomena as images.

In 1885 Hertz was in a position where he could choose between advancing to a professorship in Kiel or to move to a professorship at the Polytechnic in Karlsruhe. He chose Karlsruhe because it provided laboratory facilities. During the following four years he made the best of these facilities to make his most celebrated research: in particular, he succeeded in producing electromagnetic waves with a wavelength short enough to demonstrate that they behave like light [Buchwald, 1994]. To many physicists, including Hertz himself, this was a final proof of Maxwell's electromagnetic field theory.

Hertz became almost instantaneously world famous and was offered the prestigious chair of theoretical physics at Berlin University, as Kirchhoff's successor, but he declined it in favour of a professorship at the University of Bonn which allowed him to continue experimental research. However, Hertz was to make little use of this possibility, for instead he finished two theoretical papers on electromagnetism. The first contained a rather axiomatic treatment of Maxwell's equations in their now familiar form, while the second dealt with electromagnetism of moving bodies in a way that was soon surpassed by Hendrik Lorentz's theory of the electron (§60) and Albert Einstein's theory of relativity (§63). After a brief experimental study of cathode rays Hertz turned to theoretical mechanics.

2 MECHANICS, A RACE WITH DEATH

At the end of March 1891 Hertz wrote to Felix Klein that he had begun to think about mechanics, in particular about the theory of energy. He promised to contribute a paper

on this subject to a projected publication and anticipated that it would require between a half and a full year of work. In fact it took Hertz almost three years to finish his research. There were several reasons for this delay. First, he expanded the work from a paper on the concept of energy to a monograph containing a complete reorganization of the foundations and principles of this science. Second, his teaching, his duties as a head of a research laboratory, his duties as a physics celebrity, and finally his fatal illness prevented him from working full time on the project. Third, he found it very time consuming to write this logically tightly knit theoretical book that he sometimes characterized as his mathematical work. While he was working on it he was often depressed from the lack of progress, and he admitted in a letter of 19 May 1893: 'I often think I should not have begun it' [Fölsing, 1997, 500].

In 1892 Hertz contracted an infection of the nose, and soon it spread to other cavities of his head. Although various treatments helped temporarily, the painful disease made him unfit for work during prolonged periods. He also gradually realized that the infection could be fatal, and at several occasions he doubted that he would live to see the book on mechanics to the end. Yet, on 3 December 1893 he sent two-thirds of the manuscript off to the Barth Publishing House, with whom he had negotiated a favourable contract. According to Hertz the last third of the manuscript still needed 'a final touch'. However, his condition rapidly worsened, blood poisoning supervened, and on 1 January 1894 he died only 36 years old.

Hertz had paid his assistant Philipp Lenard to make a copy of the manuscript, and after Hertz's death Lenard saw the book through press, making a few minor final touches to the last third of the manuscript. The book appeared in the summer of 1894.

3 WHY MECHANICS?

From a modern perspective it may seem strange that a young and celebrated physicist, who had just made a decisive break through in one of the hottest areas of physics would turn to a classical subject such as mechanics. But for Hertz and his contemporaries this was a rather natural next step. With varying vigour Maxwell had suggested that the electromagnetic field should somehow be explained in mechanical terms as matter in motion. Hertz's axiomatic treatment of Maxwell's theory almost completely avoided any allusion to such a mechanical reduction, but that does not mean that he did not endorse a mechanistic reductionist program. Indeed, his opening words of the *Principles of mechanics* were: 'All physicists agree that the problem of physics consists in tracing the phenomena of nature back to the simple laws of mechanics'. Moreover, he declared that a mechanical explanation of electromagnetism 'seems to be nearly realized'. Thus, for Hertz mechanics was the fundamental discipline of physics to which all other disciplines, including electromagnetism, should ultimately be reduced. However, he shared a feeling widespread among his contemporaries that there was something rotten in the foundations of mechanics. During the decades preceding his book several critical works had appeared: in particular, Hertz owed much to Mach's *Die Mechanik in Ihrer Entwickelung historisch-kritisch dargestellt* of 1883. Several mathematicians and physicists such as William Thomson and P.G. Tait (1867, 1879–1883: see §40), Kirchhoff (1876–1877) and Carl Neumann (1888) wrote new

treatises in order to provide a more satisfactory foundation for mechanics. However, the many lectures on mechanics that Hertz had attended as a student had reinforced his feeling that the problems had not been solved in a satisfactory way.

One of the main problems that were up for debate was the role of unobservable atoms and molecules and the forces at a distance assumed to act between them. A group of positivist and phenomenologically oriented physicists and chemists believed that if physics was based on a concept of energy, one could avoid making appeal to unobservables and forces at a distance. When Hertz began his work on mechanics he had such an energetic image of nature in mind. However, he soon rejected energy as an insufficiently clear basic notion and convinced himself that physics could not do without unobservables.

Already in his 1884 lectures in Kiel Hertz had argued that field theories of mediated action could explain all the effects that were usually attributed to distance actions, and in connection with his experiments on electromagnetic waves he expressed the opinion that their most important consequence was to show that electromagnetism is not due to actions at a distance but is propagated in time through space. In a speech in 1889 he explained how gravity was now the only apparent action at a distance left in physics, and he suggested that even this force might turn out to be of a field-theoretical nature.

The contemporary opinion was that fields were to be described as mechanical states in an all-pervasive medium, the ether. Its nature was according to Hertz the 'all-important problem' of physics. However, he felt that one could only begin to understand the ether after one had removed all imperfections from the principles of mechanics. Thus, he considered his *Principles of mechanics* as a critical new foundation of mechanics necessary for the study of ordinary matter and in particular of the ether. He hoped it would eventually lead to the understanding of all interactions and ultimately of all natural processes, at least all non-living processes.

4 IMAGES OF NATURE

The contents of Hertz's book are summarised in Table 1. In the philosophical introduction he argued that we have no way of knowing how nature really works. In particular, he agreed with the positivist phenomenologists that we cannot know which unobservable elements really exist. Yet he argued, that any reasonable theory of nature must contain unobservables. So the best we can do is to make ourselves (mental) images of nature. Such an image must correspond to the nature external to our minds in such a way that 'the necessary consequents of the images in thought are always the images of the necessary consequents in nature of the things pictured' (p. 1). In other words, an image must be able to predict nature correctly. If an image satisfies this requirement Hertz called it 'correct'. There need not be any other resemblance between nature and an image of nature, and we have no way of knowing if there are other similarities between the two.

As a second requirement of an image Hertz asked that it be '(logically) permissible', a notion akin to consistency. There may be many permissible and correct images of nature. In order to chose between them Hertz required that one should prefer the most 'appropriate'. The most appropriate image is the one that is 1) the most distinct (pictures most essential relations) and 2) the simplest (containing the least number of empty or superfluous relations).

Table 1. Summary by Chapters of Hertz's book.

Chapter	**Page**	**Topics**
xii–xxxii		Preface by Hermann von Helmholtz.
Introduction	1–49	Philosophy of images and analysis of the three competing images of mechanics.
Book 1	51	*Geometry and kinematics of material systems.*
1	53	Time, space, and mass.
2	55	Positions and displacements of points and systems.
3	69	Infinitely small displacements and paths of a system of material points.
4	88	Possible and impossible displacements. Material systems.
5	100	On the paths of material systems.
6	119	On the straightest distance in holonomic systems.
7	137	Kinematics.
	153	Concluding note on Book 1.
Book 2	155	*Mechanics of material systems.*
1	157	Time, space, and mass.
2	162	The fundamental law.
3	170	Motion of free systems.
4	199	Motion of unfree systems.
5	235	Systems with concealed masses.
6	286	Discontinuous motion.
	306	Concluding note on Book 2.
	309	Index to definitions. [End 312.]

Most of the introduction deals with a comparison between three images of mechanical nature: 1) The ordinary Newtonian–Laplacian image, which operates with four basic concepts: time, space, mass and force; 2) The energetic image, which also operates with four basic concepts: time, space, mass and energy; and 3) Hertz's image, which operates with only three basic concepts: time, space and mass.

As emphasized above, Hertz had by 1890 come to the conclusion that forces acting at a distance had no place in physics; so it was a natural consequence that his mechanics had to do without them. However, he went much further in his book, for he completely excluded the concept of force (be it acting at a distance or through a medium) and the related concept of potential energy as a fundamental notion. It is obvious that if an image of nature can do without forces or potential energy, such an image must be simpler than, and thus preferable to, an image including such inessential idle wheels. This is how one might expect that Hertz would argue in favour of his own image of mechanics. However, rather than emphasizing its simplicity, he stressed its permissibility.

Hertz referred to many logical problems encountered in usual mechanics, and he concluded that the concept of force was to blame for most of them. In this situation textbooks typically added clarifying comments in order to dispel the confusion. However, he sharply pointed out that one cannot get rid of inconsistencies by adding new relations, but only by

leaving out something from the image. In particular, according to Hertz one could get rid of the inconsistencies of mechanics by leaving out the concept of force as a basic concept.

In order for different parts of a mechanical system to be able to interact, Hertz allowed what he called 'connections'. These are purely geometric relations expressible as first-order homogeneous differential equations in the coordinates. Contrary to Newton's second law, these differential equations do not involve time. This is probably why Hertz preferred connections to forces. Moreover, he admitted that we cannot give a correct image of nature without admitting other unobservables. However, contrary to the unobservables of the two other images, force and potential energy, which are of an entirely other nature than the other three basic notions of time, space and mass, Hertz's image operated with concealed mass, which was supposed to be entirely similar to ordinary mass. The only difference is that the concealed masses are not directly connected to our sensory apparatus.

Thus to Hertz a mechanical system consists of a number of ordinary mass-points connected to each other and to a system of concealed masses. He did not deal with continuum or fluid mechanics except for a passing remark to the effect that such systems could be dealt with by going to the limit.

5 GEOMETRY OF SYSTEMS OF POINTS

Most books on mechanics begin with a chapter on the motion of one point mass. Hertz on the other hand, began head on with systems. This is a natural consequence of his exclusion of forces. Indeed, there is not much one can say about the motion of one point when it cannot move in a force field. Still, Hertz treated systems of points in a way that paralleled the way one usually described one point. To that end he introduced a differential geometric formalism that he called a geometry of systems of points. A system of n point masses is described by the rectangular coordinates and the mass of each of its point masses. Let $x_{3\mu-2}, x_{3\mu-1}, x_{3\mu}$ denote the rectangular coordinates of the μth point mass and let $m_{3\mu-2} = m_{3\mu-1} = m_{3\mu}$ denote its mass. If the system is displaced such that the coordinate x_μ is increased by the value dx_μ then Hertz defined the length of the displacement ds by

$$ds^2 = \frac{1}{m}\sum_{\mu=1}^{3n} m_\mu\, dx_\mu^2, \tag{1}$$

where m is the total mass of the system. Hertz also described the system by generalized coordinates, that is a set of parameters $q_1, q_2, \ldots, q_r$, which completely determines the configuration of the system. In terms of such generalized coordinates the line element (1) will be expressed as a more general positive quadratic form

$$ds^2 = \sum_{\rho=1}^{r}\sum_{\sigma=1}^{r} a_{\rho\sigma}\, dq_\rho\, dq_\sigma. \tag{2}$$

Hertz's geometry of systems of points is a Riemannian geometry of configuration space with the metric defined by (1) or (2). Hertz's colleague in Bonn, Rudolf Lipschitz, had already in 1872 suggested an even more far-reaching geometrization of mechanics, but

Hertz was the first physicist who used such a geometric formalism that allowed him to describe a system as one point in a higher dimensional space.

Hertz continued to define the angle (s, s') between two displacements ds and ds' by the formula

$$m\,ds\,ds'\cos(s,s') = \sum_{\mu=1}^{3n} m_\mu\,dx_\mu\,dx'_\mu, \tag{3}$$

or in generalized coordinates

$$m\,ds\,ds'\cos(s,s') = \sum_{\rho=1}^{r}\sum_{\sigma=1}^{r} a_{\rho\sigma}\,dq_\rho\,dq'_\sigma. \tag{4}$$

The notion of angle allowed him to define the curvature of a path of the system by

$$c = \frac{d\varepsilon}{ds}, \tag{5}$$

where $d\varepsilon$ denotes the angle between the directions of the path at the beginning and at the end of a path element ds.

Hertz supposed that the different point masses of a mechanical system are related through 'connections' that can be expressed in the form of first-order homogeneous differential equations in the coordinates of the system

$$\sum_{\nu=1}^{3n} X_{\iota\nu}dx_\nu = 0, \quad \iota = 1, 2, \ldots, i, \tag{6}$$

or in generalized coordinates

$$\sum_{\rho=1}^{r} q_{\chi\rho}\,dq_\rho = 0, \quad \chi = 1, 2, \ldots, k, \tag{7}$$

where $X_{\iota\nu}$ and $q_{\chi\rho}$ are functions of $x_1, x_2, \ldots, x_{3n}$ and $q_1, q_2, \ldots, q_r$ respectively. This system of differential equations may be integrable such that the connections can be expressed in integral form

$$F_\iota(x_1, x_2, \ldots, x_{3n}) = C_\iota, \quad \iota = 1, 2, \ldots, i, \tag{8}$$

or

$$F_\chi(q_1, q_2, \ldots, q_r) = C_\chi, \quad \chi = 1, 2, \ldots, k. \tag{9}$$

In such cases Hertz named the system 'holonomic'. In a holonomic system it is possible to reduce the number of coordinates to $3n - i\,(r - k)$ free coordinates, that is, to coordinates that are not constrained by any of the equations (6)–(9).

The motion of a rolling ball can be described by connections like (6) and (7), but these differential equations cannot be integrated in the form (8) and (9). In order to allow for

such motions Hertz allowed non-holonomic systems. He was not the first to make the distinction between holonomic and non-holonomic systems, but he was the first to suggest the name 'holonomic', still in use to day. Moreover, his lucid analysis of the problems one encounters when dealing with non-holonomic systems made his book the origin of much of the subsequent work on such systems.

A displacement ds of a system was called 'possible' if it satisfies the connections, and a path is possible if it consists of possible displacements. Among all possible paths of a system Hertz singled out the 'straightest' (arts. 151–154). It consists of straightest line elements ds, i.e. line elements that have a smaller curvature (5) than all other possible line elements with the same starting point and starting direction. Using Lagrange's method of multipliers Hertz derived the differential equations

$$\sum_{\sigma=1}^{r} q_{\chi\rho} q_{\sigma}'' + \sum_{\sigma=1}^{r}\sum_{\tau=1}^{r}\left(\frac{\partial a_{\rho\sigma}}{\partial q_\tau} - \frac{1}{2}\frac{\partial a_{\sigma\tau}}{\partial q_\rho}\right) q_\rho' q_\tau' + \sum_{\chi=1}^{k} q_{\chi\rho}\Pi_\chi = 0, \quad \rho = 1, 2, \ldots, r. \tag{10}$$

Here Π_χ are Lagrangean multipliers and '$'$' denotes differentiation with respect to the curve length along the path. This equation combined with the equation (7) of connection gives a system of second-order differential equations for the straightest path.

Towards the end of the first book Hertz introduced the concept of time and also kinematic concepts such as velocity, momentum and acceleration that depend on time. They are examples of what he called a 'vector quantity', that is 'any quantity which bears a relation to the system and which has the same kind of mathematical manifold as a conceivable [that is, not necessarily possible] displacement of the system'. In order to conform to the usual Lagrangian and Hamiltonian formalisms of mechanics, Hertz introduced so called components of a displacement (or another vector quantity) along a coordinate. They correspond to what we now call 'covariant components' of the vector. He seems to have arrived at this notion independently of the simultaneous development of tensor calculus in the mathematics community [Reich, 1994].

Hertz also introduced the energy of a system. In his image all energy is kinetic energy. It can be expressed by the same quadratic form as ds with the differentials dx (dq) replaced by the corresponding time derivatives $\dot{x}$, $(\dot{q})$:

$$E = \frac{1}{2}\sum_{\nu=1}^{3n} m_\nu \dot{x}_\nu^2 = \frac{1}{2} m \left(\frac{\partial s}{\partial t}\right)^2 = \frac{1}{2} m \sum_{\rho=1}^{r}\sum_{\sigma=1}^{r} a_{\rho\sigma}\dot{q}_\rho \dot{q}_\sigma . \tag{11}$$

In terms of the energy, the component of the momentum p_ρ of the system along the coordinate q_ρ can be expressed by

$$p_\rho = \frac{\partial E}{\partial \dot{q}}. \tag{12}$$

This is the usual Lagrangean definition of the generalized momentum, and it explains why Hertz defined the components along a coordinate in the way that he did.

6 DYNAMICS

At the beginning of the second Book on dynamics, Hertz introduced his only law of motion. It was intentionally formulated in conformity with Newton's first law: '*Fundamental Law*. Every free system persists in its state of rest or of uniform motion [that is, with constant speed ds/dt] along a straightest path'.

Introducing time t as the independent variable into the equation (10) of the straightest path, and using that $v = ds/dt$ is a constant, Hertz deduced the following equation of motion for a free system:

$$m\left[\sum_{\sigma=1}^{r} a_{\rho\sigma}\ddot{q}_\sigma + \sum_{\sigma=1}^{r}\sum_{\tau=1}^{r}\left(\frac{\partial a_{\rho\sigma}}{\partial q_\tau} - \frac{1}{2}\frac{\partial q_{\sigma\tau}}{\partial q_\rho}\right)\dot{q}_\sigma\dot{q}_\tau\right] + \sum_{\chi=1}^{k} q_{\chi\rho}Q_\chi = 0, \quad \rho = 1, 2, \ldots, r. \tag{13}$$

Here for abbreviation he set $mv^2\Pi_\chi = Q_\chi$. From this equation of motion he could derive Lagrange's equations, which in the case of a holonomic system can be written

$$\frac{d}{dt}\left(\frac{\partial E}{\partial \dot{q}_\rho}\right) - \frac{\partial E}{\partial q_\rho} = 0, \tag{14}$$

where the q_ρs are free coordinates. Further he could deduce Hamilton's equations, D'Alembert's principle, energy conservation, and many of the other well known principles of mechanics.

In many cases (for example, the solar system) observable mechanical systems do not seem to move according to the fundamental law. Hertz called such systems 'unfree'. He postulated that every unfree system is only a part of a larger free system, whose remaining part is concealed. In such cases the problem consists in describing the motion of the observable unfree system without direct reference to the motion of the concealed subsystem. He showed how this is possible in special cases, in particular when the connections between the observable and the concealed parts of the system can be expressed as the sharing of a generalized coordinate. In this case he defined the concept of the force impressed by the concealed system on the observable system. Its component along a shared coordinate is equal to the Lagrange multiplier Q_χ in (13) to which the coupling gives rise. In this way Hertz was led to the same equations of motion (for example, d'Alembert's principle) for the unfree system as in ordinary mechanics. However, where the forces entering into the equations of ordinary mechanics are basic quantities, they are only derived quantities in Hertz's mechanics, resulting from the coupling of the unfree system with a concealed system. And where one of these equations such as d'Alembert's principle is taken as a basic law of motion in ordinary mechanics, every one of them is in Hertz's mechanics mathematical consequences of the simpler fundamental law.

Finally, Hertz defined a special type of unfree systems acted on by forces, the so-called conservative systems. The concealed subsystem of such a system have some coordinates, called the parameters, that are shared with the observable system. They give rise to the coupling. All the other coordinates of the concealed system are assumed to be cyclic coordinates, that is coordinates that do not enter explicitly into the expression of the line element ds, or equivalently into the expression of the energy; the corresponding velocities

q may, or rather must, enter into these expressions. Moreover the energy of the concealed system is assumed to be approximated sufficiently well by a quadratic form in the cyclic velocities. Now, if the observable and the concealed system interact at all there are non-cyclic coordinates (parameters) of the concealed system and in that case the non-cyclic velocities must enter into the quadratic form expressing the energy of the concealed system. However, if the cyclic velocities of the concealed system are much larger (and the masses of the concealed system is much smaller) than the velocities (and masses) of the observable system, the velocities corresponding to the parameters only give a small contribution to the energy of the concealed system. Thus, in that case the system is conservative in Hertz's sense.

In a conservative system Hertz defined the potential energy of the observable unfree part as the (kinetic) energy of the concealed part. Therefore the sum of the kinetic and the potential energy is conserved and Hertz could derive the usual formulation of Lagrange's and Hamilton's equations for conservative systems. Moreover, for holonomic conservative systems he could derive the usual integral variational principles such as the principle of least action and Hamilton's principle and the entire Hamilton formalism. For non-holonomic systems, however, Hertz emphasized and demonstrated that the integral variational principles fail when they are formulated in their natural form. A few years later, in 1896, Otto Hölder showed that a slight reformulation can save the principles even for non-holonomic systems.

Hertz was not the first to utilise the special properties of cyclic coordinates. In particular E.J. Routh had shown in 1877 that one can ignore cyclic coordinates if one replaces the ordinary Lagrangian function with a modified Lagrangean. This corresponds to the introduction of apparent forces or apparent potential energies. The same idea was used by Helmholtz in two papers on thermodynamics from 1884 and 1886, in which he introduced the idea of adiabatic cyclic systems, which provided Hertz with the technical basis for his treatment of forces and potential energy. Even before Hertz, J.J. Thomson had suggested that one should be able to explain all forces as the result of ignoring cyclic coordinates, or as he put it: 'From this point of view all energy is kinetic' [Thomson, 1888, 14].

This was precisely Hertz's standpoint. However Hertz declared that he only heard of Thomson's radical standpoint at a late stage in his work on mechanics. Yet, one might ask: what was new in Hertz's image of mechanics as compared with the ideas put forward by Helmholtz and Thomson? The main novelty was that Hertz developed mechanics from scratch without using the concept of force. Helmholtz and Thomson, on the other hand, had appealed to the usual formulation of mechanics that introduced forces at the outset. 'I endeavour from the start to keep the elements of mechanics free from that which von Helmholtz only removes by subsequent restriction from the mechanics previously developed.' he wrote in his preface.

Hertz's book gave a new and mathematically very clear deduction of the principles of mechanics from a new minimal system of assumptions. In this way he was also able to clarify the mutual logical connection between the different principles. In particular, he insisted that his method showed how a priori and empirical elements entered into the various principles. The concept and properties of time and space were according to him, a priori intuitions in Immanuel Kant's sense. The fundamental law of motion, on the other hand

was empirical in nature, and according to him the only empirical element of his mechanics. Thus, for him the question of correctness (in the sense of his image theory) of his image of nature was simply reducible to the question of the correctness of this one law. Of course history proved his analysis wrong. Indeed, only 11 years later Einstein and Hermann Minkowski changed our understanding of time and space, that is precisely the elements of mechanics that according to Hertz were a priori and unalterable (compare §63).

7 RECEPTION AND IMPACT

The reviews of Hertz's *Mechanics* are listed in [Baird et alii, 1998, 284]. When reading them one must keep in mind that in addition to being reviews of a book they were often formed as eulogies of its recently deceased author who most physicists had considered as Helmholtz's natural successor as the leader of the German physics community. This probably made the reviews more positive than they would otherwise have been. Criticisms were often formulated as questions that the world was now unable to ask the author. The following quotation from Boltzmann is typical: Hertz 'created a strikingly simple system of mechanics based on very few but to be sure logically quite natural principles. Regrettably, at the same moment his voice fell silent forever, leaving unanswered all the thousand open questions that surely I am not the only one to have on my mind' [Boltzmann, 1900, 84].

The reactions to Hertz's mechanics were broadly the same. Most reviews listed the following merits: Its philosophical sophistication, its rigorous mathematical structure, its avoidance of forces, and its intuitively pleasing formulation of the fundamental law of motion. As its main weakness most reviews mentioned its complete neglect of the problem of how to account for the actual motion of even simple systems in nature, such as those that the usual image of mechanics describes by way of forces. In Hertz's image this problem boils down to the problem of constructing a concealed system and a system of connections to the observable system, such that the fundamental law applied to the total system will give the observable part a motion that correspond to its observed motion in nature. To Hertz this question was probably equivalent to the question of the nature of the ether. The book was supposed to clarify the basis for this question, but Hertz explicitly reserved a discussion of the problem of the ether itself to later experimental work. Still, it is natural that reviewers asked themselves this question and were disappointed not to find the answer in Hertz's book.

Most reviewers, including Helmholtz, who wrote a preface to Hertz's book, suspected that if it was at all possible to find a concealed system that would account for the observed motions of natural systems, it would be so terribly complicated that it would be hard to argue that this image of the world was simpler than the traditional image. G.F. FitzGerald even wondered how it would be possible to avoid entanglements of the connections of the system. Moreover, he criticized Hertz's use of 'space of multiple dimensions' since 'this represents the real by the unattainable' [FitzGerald, 1895]. Philosophically minded physicists and philosophers such as Mach, Pierre Duhem and Henri Poincaré criticized Hertz for his hypothesis about the existence of concealed masses. Still, Mach praised Hertz's presentation of mechanics as the one that best lived up to his ideals. Duhem, on the other hand, regarded Hertz's book as the last step in a series of misconceived British mechanistic

explanations of physics. 'Hertz's mechanics is less of a doctrine than a project or a program of a doctrine'. When Duhem wrote this in 1903 many philosophers and physisists had abandoned the mechanistic reductionist program either for an electromagnetic world view, or for an energetic or phenomenological approach.

Still, several physicists, in particular Boltzmann, a late adherer to the mechanistic world view, encouraged the physics community to pursue Hertz's 'programme for the future'. He discussed technical details with the mathematician Alexander Brill, and instructed his student Paul Ehrenfest to write his thesis on the motion of rigid bodies in a fluid from a Hertzian point of view. Moreover, in 1916 F.X. Paulus in Vienna showed how one can construct concealed systems that could account for simple forces [Paulus, 1916].

But by and large Hertz's program was followed by very few. The reason is to be found in the completely new turns that physics took about a decade after Hertz's book appeared. They made his mechanics appear more as a brilliant conclusion of an era of classical physics, rather than as a program for future research in physics.

Yet Hertz's book had substantial influence on the development of science and philosophy. The use of differential geometry was soon transferred to ordinary mechanics by Lorentz in 1902, and it had a profound influence on the advanced presentations of mechanics later in the 20th century. Moreover, Hertz's clear distinction between observable nature and the theories (images) we make of it probably facilitated the highly abstract formalism of quantum mechanics with its sharp distinction between the formalism and the observable consequences. Finally, Hertz's philosophical introduction was a source of inspiration for philosophers, in particular Ludwig Wittgenstein [Barker, 1980], and for scientists. For example, David Hilbert conceived of his *Grundlagen der Geometrie* of 1899 (§55) as a geometric parallel to Hertz's clear development of the foundations of mechanics, and he considered his own requirements of an axiomatic system, consistency, completeness and independence, to correspond to Hertz's requirements of an image: permissibility, correctness and simplicity [Corry, 1997].

BIBLIOGRAPHY

Baird, D., Hughes, R.I.G. and Nordmann, A. (eds.) 1998. *Heinrich Hertz*: *classical physicist, modern philosopher*, Dordrecht: Kluwer.

Barker, P. 1980. 'Hertz and Wittgenstein', *Studies in history and philosophy of science, 11*, 243–256.

Boltzmann, L. 1900. 'Über die Entwicklung der Methoden der theoretischen Physik in neuerer Zeit', *Jahresbericht der Deutschen Mathematiker-Vereinigung, 8*, 71–95.

Buchwald, J. 1994. *The creation of scientific effects*: *Heinrich Hertz and electric waves*, Chicago: University of Chicago Press.

Corry, L. 1997. 'David Hilbert and the axiomatization of physics (1894–1905)', *Archive for history of exact sciences, 51*, 83–198.

FitzGerald, G.F. 1895. Review of Hertz's *Die Prinzipien der Mechanik, Nature, 51*, 283–285.

Fölsing, A. 1997. *Heinrich Hertz. Eine Biographie*, Hamburg: Hoffmann und Campe.

Hertz, H. 1892. *Untersuchungen ueber die Ausbreitung der elektrischen Kraft*, Leipzig: Barth. [English trans.: *Electric waves*, London: Macmillan, 1900.]

Hertz, H. 1894–1910, *Gesammelte Werke*, 3 vols., Leipzig: Teubner.

Hertz, H. 1999. *Die Constitution der Materie* (ed. A. Fölsing), Berlin: Springer.

Hertz, J. (ed.) 1977. *Heinrich Hertz. Erinnerungen, Briefe, Tagebücher/Memoirs, letters, diaries*, 2nd ed., Weinheim: Physik Verlag; San Francisco: San Francisco Press.

Klein, M.J. 1970. *Paul Ehrenfest*, vol. 1, *The making of a theoretical physicist*, Amsterdam: North Holland.

Lorentz, H.A. 1902. 'Eenige beschouwingen over de Grondstellingen der Mechanica, naar Anleiding dan "Die Prinzipien der Mechanik" van Hertz', *Amsterdam Zittingsverlag Academie voor Wetenschappen, 10*, 876–895. [Repr. in *Abhandlungen über theoretische Physik*, vol. 1 (1907), 1–22. English trans.: 'Some considerations on the principles of dynamics, in connection with Hertz's "Prinzipien der Mechanik" ', in *Amsterdam Proceedings*, (1901–1902), 713–732.]

Lützen, J. 1999. 'Geometrizing configurations. Heinrich Hertz and his mathematical predecessors', in J.J. Gray (ed.), *The symbolic universe. Geometry and physics 1890–1930*, Oxford: Oxford University Press, 25–46.

Lützen, J. 2005. *Mechanistic images in geometric form*: *Heinrich Hertz's Principles of Mechanics*, Oxford: Oxford University Press.

Paulus, F.X. 1916. 'Ergänzungen und Beispiele zur Mechanik von Hertz', *Sitzungsberichte der Academie der Wissenschaften zu Wien, mathematisch-physikalische Klasse, (2a) 125*, 835–882.

Reich, K. 1994. *Die Entwicklung des Tensorkalküls*, Basel: Birkhäuser.

Thomson, J.J. 1888. *Applications of dynamics to physics and chemistry*, London: Macmillan. [Repr. New York: Dover, 1968.]

CHAPTER 53

HEINRICH WEBER, *LEHRBUCH DER ALGEBRA* (1895–1896)

Leo Corry

As the last important textbook on algebra published in the 19th century, Weber's *Lehrbuch* presents a faithful image of algebraic knowledge as then conceived. Although many of the abstract concepts that became central to the structural conception of algebra after 1930 were well known to Weber, they play a relatively secondary role here: algebra was still the discipline of polynomial equations and polynomial forms.

First edition. 2 vols., Braunschweig: Vieweg, 1895–1896. 772 + 876 pages. Vol. 3 published separately as *Elliptische Funktionen und algebraische Zahlen*, 1891. 764 pages.

Second edition. Vols. 1–3: Braunschweig: Vieweg, 1898–1908.

Reduced edition in one volume. Braunschweig: Vieweg, 1912. [Photorepr. New York: Chelsea, 1961.]

French translation of the 2nd edition. *Traité d'algèbre supérieure* (trans. J. Griess), Paris: Gauthier–Villars, 1898.

Related articles: Dirichlet (§37), Hilbert on number theory (§54), van der Waerden (§70).

1 BACKGROUND

The discipline of algebra underwent significant changes between the last third of the 19th century and the first third of the 20th century. They comprised the addition of important new results, new concepts and new techniques, as well as meaningful changes in the way that the very aims and the scope of the discipline were conceived by its practitioners. Over the 19th century algebraic research had meant mainly research on the theory of polynomial equations and the theory of polynomial forms, including algebraic invariants. The ideas implied by Evariste Galois's works became increasingly visible and central after their publication by Joseph Liouville in 1846. Together with important progress in the theory of

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

fields of algebraic numbers, especially in the hands of Leopold Kronecker and Richard Dedekind (§37), they gave rise to an increased interest in new concepts such as groups, fields and modules.

A very popular textbook of algebra since the middle of the century was the *Cours d'algèbre supérieure* by Joseph Serret, which underwent three editions in 1849, 1854 and 1866. In these successive editions it gradually incorporated the techniques introduced by Galois, and became the first university textbook to publish a full exposition of the theory. Still, it continued to formulate the main results of Galois theory in the traditional language of solvability dating back from the works of Joseph-Louis Lagrange and Niels Henrik Abel at the beginning of the century (§29); in doing so, it did not even include a separate discussion of the concept of group. A second contemporary textbook was Camille Jordan's *Traité des substitutions et des équations algèbriques*, which already included a more elaborate presentation of the theory of groups but still treated this theory as subsidiary to the main task of discussing solvability conditions for polynomials [Jordan, 1870].

A completely different image of algebra is embodied much later in Bartel L. van der Waerden's textbook *Moderne Algebra* [van der Waerden, 1930, 1931]. Here we are presented for the first time with a discipline at the center of which stands the general idea of an abstract algebraic structure that is instantiated in various particular species such as groups, rings and fields calling for being elucidated with the help of a standard set of tools (§70). This is the image that came to dominate research in the 20th century.

Weber's *Lehrbuch der Algebra* stands midway between these two poles. It incorporates an entire body of new individual ideas and techniques developed along the 19th century, and in doing so it provides a full picture of what algebraic knowledge looked like at the time. In spite of the knowledge it adds over books like Serret's or Jordan's, the picture of algebra it presents does not differ essentially from theirs. On the other hand, in spite of including a great deal of material that would eventually be incorporated as the basis of van der Waerden's presentation, it does not envisage the kind of fundamental change in conception that *Moderne Algebra* intended to imply.

2 HEINRICH WEBER'S CAREER

Heinrich Weber (1842–1913) studied in Heidelberg and Leipzig. He habilitated in Königsberg in 1866, and taught there until 1883, except for the years 1870 to 1875 when he was professor at the ETH Zürich. He later spent several years at the Charlottenburg Technological Institute (near Berlin) and in Marburg, and was full professor in Göttingen between 1892 and 1895. Finally he moved to Strasbourg, where he remained until his death in 1913 [Schappacher and Volkert, 1997].

Weber's Königsberg years were the most productive of his successful career. The tradition of analysis and mathematical physics developed in this university, under the leadership of Carl Gustav Jacobi, Franz Neumann and, somewhat later, Friedrich Richelot, was a main force behind the increasing dominance attained by Germany in the mathematical world over the 19th century. Weber was but one of the outstanding mathematicians whose names came to be connected with that school. The early careers of Adolf Hurwitz, Hermann Minkowski and David Hilbert were also later associated with this institution and their

works were decisively influenced by its tradition. Weber's mathematical activities spanned many different domains of mathematics such as algebra, number theory and mathematical physics. In his historical account of the development of mathematics in the 19th century, Felix Klein described Weber as the most versatile representative of that trend, of which Klein himself so proudly felt part and which sought to elaborate the interconnections between mathematical domains such as the theory of invariants, the theory of polynomial equations, the theory of functions, geometry and the theory of numbers [Klein, 1926, 275].

Weber's *Lehrbuch* was only one among several important works that he published and reached a wide mathematical audience. For instance, together with J. Wellstein he put out a widely-read *Enzyklopädie der Elementar-Mathematik* [Weber and Wellstein, 1903–1907]. He also collaborated with Richard Dedekind in two further important projects. One was a seminal article on the theory of algebraic functions [Dedekind and Weber, 1882]. The other one was the edition of Bernhard Riemann's mathematical papers, published in 1892. It is very likely that without Weber's active help, Dedekind would have never completed the edition.

When the first edition of Volume I of the *Lehrbuch der Algebra* appeared in 1895, Weber was well aware of the latest advances in algebra, and in particular of the possibility of formulating new algebraic concepts in purely abstract terms. As a matter of fact, in [Weber, 1893] he had been the first to publish abstract definitions of both groups and fields within the framework of a single article. Moreover, his research on algebraic functions in collaboration with Dedekind shows that he was deeply acquainted with the latter's theory of ideals, a theory that played a central role in the rise of the structural approach to algebra. And yet, when the time came for presenting the current state of knowledge in the discipline, he chose to present such concepts as playing only a relatively marginal role within it.

3 THE INTRODUCTION TO THE *LEHRBUCH*

In the preface to the first volume of the *Lehrbuch* Weber explained that the development of algebra over the preceding decades had rendered the existing textbooks obsolete and had brought about the need for a new coherent presentation of results and their applications. Among the books he had in mind were Serret's *Course* and Jordan's *Traité*. The aim of his first volume was to present the 'elementary parts of algebra'; namely, all that may be subsumed under the designation of 'formal algebraic manipulation' ('*Buchstabenrechnung*'), beginning with the rules for the determination of the roots of an equation and finishing with an exposition of Galois theory. Weber explicitly acknowledged Dedekind's influence in consolidating his long-standing interest in algebra. This influence acted mainly through the notes of Dedekind's Göttingen lectures of 1857–1858 on Galois theory, the manuscript of which Weber had had the opportunity to read.

The problem of finding the roots of polynomial equations dominates a considerable portion of the book. Like all previous books in algebra, the whole theory of polynomials appears here as conceptually dependent on a thorough knowledge of the properties of the various systems of numbers. The fashion in which these systems are introduced in order to provide the necessary conceptual infrastructure differs considerably, however, from previous ones in that it is strongly based on the notion of set ('*Mannigfaltigkeit oder Menge*').

Following Dedekind, Weber introduced the concept in what we would call today a naive formulation: a system of objects or elements of any kind, such that for any given object one can always say whether it belongs to the set or not. Weber also introduced additional, related concepts such as ordered sets, discrete and dense ('*dicht*') sets (exemplified by the integers and the rationals), cuts ('*Schnitte*') and continuity ('*Stetigkeit*')—all of them as previously defined by Dedekind.

Within this framework of ideas, the rational numbers are introduced as a dense but discontinuous set and the real numbers as the set of cuts of the rationals. In spite of its markedly abstract orientation, Weber's definitions of the various number systems essentially differ from what became the standard in 20th-century mathematics. He conceived these systems as well-known, specific mathematical entities whose properties, although originating in free acts of creation of the human spirit, are given once and for all in advance. In the image of algebra embodied in Weber's book, the algebraic properties of number systems do not derive from those of some more basic or underlying abstract algebraic structures. Rather it is the other way around: algebra is based on the given properties of the number systems.

The introduction of the *Lehrbuch* closes with a remark on the formal manipulation of symbols. One can distinguish two main forms of the latter: identities and equations. Algebra, wrote Weber, is the discipline whose aim is the resolution of equations. This statement is not mere lip-service to the prevailing views. The contents of the book faithfully reflected this declared central role of equations, whereas other issues, such as the study of groups, appear as conceptually subsidiary to this aim. In fact, besides groups, no other abstract algebraic concept (fields, modules, rings, etc.) is systematically investigated in the *Lehrbuch*.

4 THE THREE VOLUMES

The layout of the *Lehrbuch* is outlined in Table 1. The first volume comprises three Books. The first two deal with the classical theories of polynomial equations. Within this framework, Chapters III and IV provide interesting evidence of the attachment of Weber to 19th-century images of algebra. Thus, for instance, in Chapter III, the concept of root of an equation is discussed in terms that may be classified as 'analytic': limits, continuity, ε–δ arguments, and so on. Arguments of this kind would later be excluded from standard, structural presentations of algebra. Likewise, Chapter IV deals with 'symmetric functions' that had been used by Lagrange in his early research on solvability of polynomial equations. Later, the gradual development of Galois theory as the main tool for studying solvability of polynomial equations eventually rendered symmetric functions a rather dispensable tool, yet Weber included a treatment of them in the *Lehrbuch* as part of a tradition of which his approach to algebra was part and parcel. Under the conception of algebra characteristic of this tradition, a treatment of the theory of polynomial equations should include every particular technique devised to deal with their solvability, as he indeed did here.

In Book II one finds additional discussions that are analytic in character, and that would be excluded from later textbooks of algebra. This is the case, for instance, of the theorem of Sturm discussed in Chapter VIII. It concerns the question of how many real roots of a given polynomial equation lie between two given real numbers. This, and further similar problems, are solved with the help of derivatives and other analytical tools. Likewise, Weber

Table 1. Contents by Chapters of Weber's volumes.

Volume 1	772 pages. Introduction (pages 1–25).
Book 1	*The foundations* (pages 25–270).
I	Rational functions.
II	Determinants.
III	Roots of algebraic equations.
IV	Symmetric functions.
V	Linear transformations. Invariants.
VI	The Tschirnhaus transformation.
Book 2	*The roots* (pages 271–490).
VII	Reality of roots.
VIII	Sturm's theory.
IX	Evaluation of roots.
X	Approximate evaluation of roots.
XI	Continued fractions.
XII	The theory of roots of unity.
Book 3	*Algebraic magnitudes* (pages 491–772).
XIII	Galois theory.
XIV	Application of groups of permutations to equations.
XV	Cyclical equations.
XVI	Cyclotomy.
XVII	Algebraic solution of equations.
XVIII	Roots of metacyclic equations.
Volume 2	876 pages. Book 1 *Groups* (pages 3–162).
I	General theory of groups.
II	Abelian groups.
III	Groups of cyclotomy fields.
IV	Cubic and biquadratic Abelian fields.
V	Constitution of the general groups.
Book 2	*Linear groups* (pages 163–350).
VI	Groups of Linear substitutions.
VII	Invariants of groups.
VIII	Groups of binary linear substitutions.
IX	Polyhedric groups.
X	Groups of congruences.

Table 1. (*Continued*)

Book 3	*Applications of group theory* (pages 351–552).
XI	General theory of metacyclic equations.
XII	Inflection points in third-order curves.
XIII	Double tangents in fourth-order curves.
XIV	The general theory of fifth-degree equations.
XV	Groups of linear ternary substitutions.
XVI	The problem of forms of the group G_{168} and the theory of seventh-degree equations.
Book 4	*Algebraic numbers* (pages 553–876).
XVII	Numbers and functionals of an algebraic curve.
XVIII	Theory of algebraic fields.
XIX	Relations between a field and its divisors.
XX	Lattice of points.
XXI	Number classes.
XXII	Cyclotomic fields.
XXIII	Abelian fields and cyclotomic fields.
XXIV	Number class of cyclotomic fields.
XXV	Transcendental numbers.
Volume 3	764 pages. Book 1. *Analytical part* (pages 1–320).
I	The elliptic integrals.
II	Theta functions.
III	Transformations of theta functions.
IV	The elliptic functions.
V	The modular functions.
VI	Multiplication and division of elliptic functions.
VII	Theory of transformation equations.
VIII	The group of transformation equations and the fifth-degree equation.
Book 2	*Quadratic fields* (pages 321–412).
IX	Discriminants.
X	Algebraic numbers and forms.
XI	Ideals in quadratic fields.
XII	Rings (‘*Ordnungen*’) in quadratic fields.
XIII	Equivalence according to groups of numbers.
XIV	Composition of forms and ideals.

Table 1. (*Continued*)

XV	Signature (‘*Geschlecht*’) of quadratic forms.
XVI	Number class in quadratic fields.
Book 3	*Complex multiplication* (pages 413–562).
XVII	Elliptic functions and quadratic forms.
XVIII	Galois group of class equations.
XIX	Calculation of class invariants.
XX	The multiplication equation in the complex multiplication.
XXI	The norm of class invariants $f(\omega)$.
XXII	Cayley’s derivation of the modular functions.
Book 4	*Class fields* (pages 563–622).
XXIII	The cyclotomic field.
Book 5	*Algebraic functions* (pages 623–764).
XXIV	Algebraic functions of one variable.
XXV	Functionals.
XXVI	Numerical values of algebraic functions.
XXVII	Algebraic and Abelian differentials.

discussed well-known approximation techniques: interpolation and Newton’s method are mentioned among others in Chapter X. Chapter XI deals with roots of unity: no mention whatsoever, however, is made of their group-theoretical properties.

Galois theory is finally introduced in Book III, after nearly five hundred pages of discussion on the resolution of polynomial equations. First, a field of numbers is defined as a set of numbers closed under the four operations. Indeed, the concept is extended to fields of functions or to any set closed under the four operations of addition, multiplication, subtraction and non-zero division (pp. 491–492). However, although Weber referred here to his own article of 1893, in which he had insisted upon the potential interest involved in studying finite fields, in the *Lehrbuch*, he considered only (infinite) fields of characteristic zero. In no way did he research fields as an autonomous concept with intrinsic interest, even at the relatively elementary level that he did for groups.

Groups are mentioned for the first time as late as p. 511. But even here one does not find a general treatment of groups; this is left for later chapters. At this stage, Weber considered substitutions of one root of a function with another, substitutions that may themselves be composed to form a group, and, more specifically, a finite group (p. 513). Weber defined here a group of permutations—a concept which he used in the next chapter—and the Galois group of a given field.

Chapter XIV shows the application of groups of permutations to the theory of equations. First, Weber showed that any permutation may be decomposed into transpositions and cycles. He then defined some additional, basic concepts: subgroups (‘*Teiler*’), and the cosets (‘*Nebengruppen*’) determined by a given permutation, as well as the index of a subgroup

of permutations. He explicitly stated that the aim of this whole section was to improve our understanding of the issues dealt with in the preceding chapter (p. 529). Thus the focus of interest does not lie in the study of the properties of the group of permutations as such, but only insofar as it sheds light on the theory of equations.

In the following chapters, Weber analyzed particular cases of equations using the insights provided by the already developed theory. The exposition culminates towards the end of the book, in Chapter XVII, where the algebraic solution of equations was systematically discussed. Weber acknowledged the centrality of this problem for the contemporary development of algebra, and the important contribution of group theory to its better understanding. Thus, he wrote (p. 644):

> One of the oldest questions which the new algebra has preferentially addressed is that of the so-called algebraic solution of equations, meaning the representation of the solution of an equation through a series of radicals, or their calculation through a series of root-extractions. The theory of groups sheds much light on this question.

It is in this section that Weber proves that the alternating group is simple—a result needed for the proof of the impossibility of solving the general fifth degree equation in radicals (pp. 649–652).

A thoroughly abstract definition of group, similar to that of Weber's own 1893 article, appears only in the second volume of the *Lehrbuch*. After the basic concepts of the theory of groups were introduced in the first four chapters of the second volume in a general and abstract way, Weber stated the object of the abstract study of groups. His formulation stresses the need he felt to explain to contemporaries the meaning of the very use of abstract concepts of this kind (p. 121):

> The general definition of group leaves much in darkness concerning the nature of the concept [...]. The definition of group contains more than appears at first sight, and the number of possible groups that can be defined given the number of their elements is quite limited. The general laws concerning this question are barely known, and thus every new special group, in particular of a reduced number of elements, offers much interest and invites detailed research.

Weber also pointed out that the determination of all the possible groups for a given number of elements was still an open question. It had been recently addressed by Arthur Cayley, but only for the lowest orders.

In the following chapters, Weber discussed special instances of groups (groups of characters, groups of linear substitutions, polyhedric groups) and presented some applications of group theory, such as Galois theory, invariants, and others. The last part of the second volume deals with algebraic number theory. Following Dedekind, the fields considered are only fields of numbers, rather than abstract ones.

The third volume of the *Lehrbuch* appeared in 1908; it embodied a second edition of Weber's book on elliptic functions and algebraic numbers, first published in 1891. It dealt with the reciprocal interrelations between problems and techniques of the theory of fields of algebraic numbers and of the theory of elliptic functions. It is important to notice that

a complete description of contemporary images of algebra cannot fail to stress the importance of the connection established in this third volume between these two domains, algebra and the theory of elliptic functions. The kinds of conceptual and technical interconnections that were pursued by mathematicians like Weber during the second half of the 19th century in relation to algebraic problems cover a much broader spectrum than the later, structural image of algebra may lead us to assume. This third volume touched upon some important portions of that spectrum. The existence of these kinds of interconnections underscored the difficulties inherent to the use of one and the same term, 'algebra', to denote the disciplines known by this name in the 19th and 20th centuries.

5 IMPACT

If one considers together the ideas appearing in [Weber, 1893] and in his *Lehrbuch*, then one finds a complex picture of his conception of algebraic knowledge. It comprises elements of both classical 19th-century conceptions as well as more modern ones (compare [Corry, 1996]). The central issue of the first volume of the *Lehrbuch* was the resolution of polynomial equations, and its presentation remains similar to those of those appearing in earlier textbooks of algebra. All the concepts and techniques related to Galois theory (in particular, the concepts of group and field) are introduced, to a large extent, only as ancillary to that central issue. By the end of the century, group theory was the paradigm of an abstractly developed theory, if there was any. Research on groups had increasingly focused on questions that we recognize today as structural, and, at the same time, the possibility of defining the concept abstractly had been increasingly acknowledged. More importantly, the idea that two isomorphic groups are in essence one and the same mathematical construct had been increasingly adopted: Weber [1893] exemplifies clearly this trend. Yet in his book group theory plays a role that, at most, may be described as ambiguous regarding the overall picture of algebra. For, although in its second volume, the theory of groups is indeed presented as a mathematical domain of intrinsic interest for research and many techniques and problems are presented in an up-to-date, structurally-oriented fashion, the theory appears in the first volume as no more than a tool of the theory of equations (albeit, it is now clear, a central one). Weber's book, and much more so his 1893 article, bring to the fore the interplay between groups and fields abstractly considered more than any former, similar work. However, in spite of this, the classical conceptual hierarchy that viewed algebra as based on the essential properties of the number systems is not called into question in any of these two works.

Weber's *Lehrbuch* became the standard German textbook on algebra and underwent several reprints. Its influence can be easily detected, among others, through the widespread adoption of a large portion of the terminology introduced in it. But not all of his terms were widely adopted. Thus for instance we find in his book the term 'metacyclic' groups (vol. 1, 646), which denoted the group of an equation that can be fully solvable by radicals, or 'Ordnung' (following Dedekind) to denote a ring of algebraic numbers.

At any rate, the image of algebra conveyed by Weber's book was to dominate the algebraic scene for almost 30 years, until van der Waerden's introduction of the new, structural image of algebra. But obviously, influential as the latter was on the further development

of algebra, it did not immediately obliterate Weber's influence, which can still be traced to around 1930 and perhaps even beyond. One can notice this by looking at several books published in the 1920s, such as Leonard Eugene Dickson's *Modern algebraic theories* [Dickson, 1926] and Helmut Hasse's *Höhere Algebra* [Hasse, 1926]. But the clearest sign long-standing influence of the *Lehrbuch* on algebraic activity, especially within Germany, is provided by the publication in 1924 of another textbook by Robert Fricke. He wrote it upon the request of Weber's publisher in Braunschweig, F. Vieweg, after the *Lehrbuch* had sold out. In spite of the relatively long time since the original publication, and the many important advances in algebraic research since then, Fricke chose to essentially abide by the conception of algebra embodied in Weber's presentation. He stressed this very clearly in the name he chose for his own textbook: *Lehrbuch der Algebra—verfasst mit Benutzung vom Heinrich Webers gleichnamigem Buche*.

BIBLIOGRAPHY

Corry, L. 1996. *Modern algebra and the rise of mathematical structures*, Basel: Birkhäuser (second revised edition; 2003).

Dedekind, R. and Weber, H. 1882. 'Theorie der algebraischen Funktionen einer Veränderlichen', *Journal für die reine und angewandte Mathematik, 92*, 181–290. [Repr. in Dedekind, *Gesammelte mathematische Werke*, vol. 1, 238–350.]

Dickson, L.E. 1926. *Modern algebraic theories*, Chicago: Benjamin H. Sanborn.

Hasse, H. 1926. *Höhere Algebra*, Berlin: Sammlung Göschen.

Jordan, C. 1870. *Traité des substitutions et des équations algèbriques*, Paris: Gauthier–Villars.

Klein, F. 1926, 1927. *Vorlesungen über die Entwicklung der Mathematik im 19. Jahrhundert*, 2 vols. (ed. R. Courant and O. Neugebauer), Berlin: Springer. [Repr. New York: Chelsea, 1948.]

Schappacher, N. and Volkert, K. 1997. 'Heinrich Weber; un mathématicien à Strasbourg, 1895–1913', *L'Ouvert* (Journal de l'A.P.M.E.P. d'Alsace et de l'I.R.E.M. de Strasbourg), *89*, 1–18.

Serret, J. 1849. *Cours d'algèbre supérieure*, Paris: Gauthier–Villars.

Van der Waerden, B.L. 1930, 1931. *Moderne Algebra*, 1st ed., 2 vols., Berlin: Springer.

Weber, H. 1893. 'Die allgemeinen Grundlagen der Galoisschen Gleichungstheorie', *Mathematische Annalen*, *43*, 521–549.

Weber, H. and Wellstein, J. 1903–1907. *Enzyklopädie der Elementar-Mathematik*, 1st ed., 3 vols., Leipzig: Teubner.

CHAPTER 54

DAVID HILBERT, REPORT ON ALGEBRAIC NUMBER FIELDS (‘*ZAHLBERICHT*’) (1897)

Norbert Schappacher

In this report Hilbert summed up the current state of knowledge in algebraic number theory, at the same time enriching and organising the subject in ways that were to influence developments for decades. However, the reception of the work has been somewhat mixed.

First publication. ‘Die Theorie der algebraischen Zahlkörper’, *Jahresbericht der Deutschen Mathematiker-Vereinigung, 4* (1897), 175–546.

Later edition. In Hilbert, *Gesammelte Abhandlungen*, vol. 1, Berlin and Heidelberg: Springer, 1932 (repr. 1970), 63–363. [Some modernised spelling, errata worked into the text; further corrections include some indicated in the copy of the *Jahresbericht* that Olga Taussky-Todd used at the Technical University Vienna when working on the *Gesammelte Abhandlungen*. – Thanks to C. Binder for pointing this out.]

French translation. By M.A. Levy as ‘Théorie des corps de nombres algébriques’, *Annales de la Faculté des Sciences de l’Université de Toulouse* (1909: publ. 1910), 3rd. fasc.

English translation. The theory of algebraic number fields (trans. I.T. Adamson, intro. by F. Lemmermeyer and N. Schappacher), Berlin: Springer, 1998.

Manuscript. None exists, but Hilbert’s personal copy with a few annotations is held in his *Nachlass* (Göttingen University Library Archives).

Related articles: Gauss (§22), Dirichlet (§37), Weber (§53), van der Waerden (§70).

1 A REPORT AND ALMOST A TEXTBOOK

This report by David Hilbert (1862–1943) was his first major writing after moving in 1895 to Göttingen University from the university in his home town of Königsberg. At Göttingen, he soon built up a reputation as the leading mathematician of his generation, with massive contributions to various mathematical disciplines; three others are discussed in this volume

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

(§55, §57 and §77). He also gave highly influential lecture courses, including in physics from the 1900s, and directed a number of doctoral students which was then unprecedented for a mathematician.

Hilbert’s so-called ‘Zahlbericht’ of 1897 was one of the reports on the state of mathematical disciplines commissioned by the *Deutsche Mathematiker-Vereinigung* (hereafter, ‘DMV’) which was founded in 1890, during the first years of its existence; the first ten volumes of the *Jahresbericht der DMV* contain thirteen such reports. Hilbert and Hermann Minkowski (1864–1909) were asked on the occasion of the DMV meeting at Munich in September 1893 to write a joint report covering all of number theory. They decided to divide up the work, leaving to Minkowski subjects like continued fractions, quadratic forms, and the geometry of numbers. Both started working on the report in 1894. In the end, only Hilbert’s part was completed, on 10 April 1897, but Minkowski did comment on Hilbert’s manuscript and read the galley proofs.

Unlike most of the other reports commissioned by the DMV, Hilbert’s *Zahlbericht* goes beyond the mere business of stocktaking. It gave a remarkably systematic and lucid treatment of algebraic number theory, thereby firmly establishing this discipline as a major domain of pure mathematics and providing at the same time its principal reference book for more than twenty years after its appearance, and leaving its mark on textbooks in this area until today. Already in a letter of 31 March 1896 Minkowski had predicted that the report would ‘certainly be greeted by general applause, and will push Dedekind’s and Kronecker’s works very much to the background’ ([Minkowski, 1973, 80]: compare §53). But it should be noted that Hilbert’s own most far-reaching number-theoretic works, where he envisaged general class field theory while studying nothing but the arithmetic of quadratic extensions, appeared only after the *Zahlbericht*, in 1899 and 1902.

The advanced character of the report made it obviously inaccessible for a broader readership. Hilbert taught a course in the winter of 1897–1898 where he emphasized quadratic number fields, and he subsequently encouraged Julius Sommer [Blumenthal, 1935, 398], who had followed these lectures, to write a textbook which dwells on quadratic and cubic fields as an introduction to algebraic number theory [Sommer, 1907]. Similarly, his American doctoral student L.W. Reid (1899 thesis on class number tables for cubic fields) published a strongly example-oriented textbook treating exclusively quadratic extensions [Reid, 1910]. Hilbert contributed to it an introduction where one reads: ‘The theory of numbers is independent of the change of fashion and in it one does not see, as is often the case in other departments of knowledge, one conception or method at one time given undue preeminence, at another suffering undeserved neglect’. We will briefly discuss in section 3 below how Hilbert actually chose among the various ‘conceptions and methods’ that existed in the literature on which he had to report.

2 THE PREFACE: NUMBER THEORY AND ARITHMETISATION

One reason for its great impact, apart from its striking expositional quality, was the fact that Hilbert was able to present current (algebraic) number theory as a leading mathematical discipline in tune with what he saw as the dominating values of the time. In his strong, sweeping preface, he not only recapitulates that number theory through its very origin is

marked by the 'simplicity of its foundations, the precision of its concepts, and the purity of its truths', but also lists many interrelations of number theory with various other branches of mathematics, claiming in the end that 'if I am not mistaken, the whole modern development of pure mathematics takes place principally under the badge ('*Zeichen*') of number'. And Hilbert alludes explicitly to the 'arithmetisation' of function theory by Richard Dedekind (1831–1916), Karl Weierstrass and Georg Cantor (§46, §47), and to studies in the axiomatisation of geometry, of which he was soon going to be the champion himself with his 1899 essay on the foundations of geometry (§55).

Hilbert did *not* here allude to Leopold Kronecker (1823–1891), who had been the first to suggest in print (in 1887) a programme of explicitly 'arithmetizing' all of pure mathematics, but with the exclusion of geometry and mechanics [Boniface and Schappacher, 2001, intro.], thereby implying a separation of number theory and analysis from geometry. Hilbert had still faithfully echoed this separation in his 1891 lectures on projective geometry [Toepell, 1986, 21], a separation which Dedekind shared as well, and which can even be considered as being handed down from Gauss. However, in the preface to the *Zahlbericht* he emphasized the similarity of all mathematical disciplines once they are treated 'with that rigour and completeness [. . .] which is actually necessary'.

As to the style of the *Zahlbericht*, it is meant to reflect the mature state of the theory of algebraic number fields. Hilbert tried to avoid Ernst Kummer's 'formidable computational apparatus, so that here too Riemann's principle be realised according to which the proofs ought to be forced not by calculations, but by pure thought'.

Kronecker's programme of arithmetisation had also been inspired by the desire to have number theory and its genuine methods—which, for Kronecker, were thought to be found essentially in C.F. Gauss—govern pure mathematics. Likewise Hilbert's report, in its own way, consciously and successfully portrays (algebraic) number theory as a model theory for pure mathematics, both in content and in form. It not only came out different in style from all the other reports commissioned by the DMV, but effectively created a new special type of technical mathematical treatise, marked by the exceedingly stringent overall logical organization of virtually all 19th-century literature in algebraic number theory. Hilbert delicately differentiated between *Hilfssätze* (only of momentary importance in the argument at hand), continuously numbered *Sätze*, and *Sätze* whose statements were printed in italics and were supposed to be major starting-points for future developments. With these distinctive literary features, the *Zahlbericht* echoes, from the turn of the 20th century, the role played by Gauss's *Disquisitiones arithmeticae* 96 years earlier (§22). To be sure, the mathematical-historical context in 1897 was very different from the one that Gauss's book had changed so profoundly in 1801, and the very theory of integers in an arbitrary algebraic number field, which constitutes the subject of the *Zahlbericht*, is entirely a creation of the 19th century. Yet, both works represent, each one in its time, major inthronisation rites performed by number theorists for the 'Queen of Mathematics' before the eyes of their mathematical colleagues.

3 DEDEKIND VERSUS KRONECKER, ARITHMETIC VERSUS ALGEBRA

There are several features of the *Zahlbericht* which mark the time when it was written and which may surprise the unsuspecting modern reader. In the 1860s and 1870s, algebraic

number theory had been the fairly solitary domain of research of a few individuals, among whom Dedekind in Braunschweig and Kronecker in Berlin stood out as the most visible and influential.

An alternative, completely viable and general approach by Egor Ivanovitch Zolotarev (1847–1878)—his second proposal for an algebraic number theory—was published only in 1880, after the author’s death, and was incorrectly thought by both Dedekind and Kronecker to yield as incomplete a theory as Zolotarev’s first proposal from his 1874 Russian thesis. This, added to the fact that Zolotarev had been an outsider to the German arithmetical community, was probably why Hilbert did not even mention Zolotarev in his references.

Dedekind developed his ideal-theoretic approach in three subsequent editions (1871–1894) of supplements 10 (or 11) of his edition of Dirichlet’s lectures on number theory (§37); it was to become one of the major sources of inspirations for the theory of commutative rings by Emmy Noether (1882–1935) in the 1920s. Kronecker is known to have thought about general algebraic number theory as of the 1850s, and he finally published an extensive account of his attempt at a unified theory for both algebraic number theory and the arithmetic theory of algebraic functions in one or several variables in 1882 [Kronecker, 1882]. This publication also contains numerous hints at the evolution of his ideas, especially in the case of number fields, and their relations to other authors.

Then, in the 1880s and 1890s, energetic younger people were entering the subject—on the one hand Kronecker’s pupil Kurt Hensel, and on the other hand Adolf Hurwitz (1859–1919) and Hilbert, both of whom cared little about the methodological preferences of either Dedekind or Kronecker in this area of research. According to Otto Blumenthal, Hilbert told later that once he and Hurwitz went for a walk in Königsberg where ‘one of us presented Kronecker’s proof for the unique decomposition into prime ideals, the other Dedekind’s, and we would find both awful’ [Blumenthal, 1935, 397]. In several papers of the mid 1890s, while using Dedekind’s notions of (number) field and ideal, Hurwitz defined ideals via finite sets of generators, and used a basically Kroneckerian approach via polynomials in several unknowns to derive the unique decomposition of ideals into prime ideals. This was much to Dedekind’s chagrin, who criticized this approach—which he had actually tried and developed himself earlier—as lacking methodological and conceptual purity [Dedekind, 1895]. Hilbert also published on this circle of ideas in 1894, giving a certain priority to Galois number fields; see our comments on Part 2 of the report in the next section.

In the *Zahlbericht*, ideals are defined in Dedekind’s style as sets of algebraic integers which are closed under linear combinations with algebraic integer coefficients (art. 4). But both for the uniqueness of decomposition into prime ideals in arbitrary number fields (arts. 5–6), and for the proof that the ramified primes are precisely the divisors of the discriminant (arts. 10–13), Hilbert adopts essentially the Kronecker–Hurwitz method and mentions Dedekind’s approach only in a reference.

In her comment of 1930 made for Dedekind’s *Gesammelte Werke* [Dedekind, 1895, 58], Noether strongly endorsed Dedekind’s criticism of Hurwitz, and she pointed out how long it had taken Dedekind’s point of view to enter standard courses and textbooks. She did not mention Hilbert’s *Zahlbericht* there, but Olga Taussky-Todd later remembered her criticising it, and claiming that Emil Artin, too, had accused Hilbert of having ‘delayed

the development of algebraic number theory by decades'. This may very well have been directed at the non-Dedekindian features of the text [Brewer and Smith, 1981, 82, 90].

More generally, Hilbert's *Zahlbericht* makes even less use of unifying notions from abstract algebra than one might have expected from a text written in the last decade of the 19th century. Thus, while the notion of (number) fields and their arithmetic is at the very heart of Hilbert's concept of algebraic number theory, and even though Hilbert does use the word '(Zahl)ring' for orders in algebraic number fields, this does not mean that he employs here parts of our current algebraic terminology; rather than referring to a general algebraic structure, the word 'ring' is used for certain sets of algebraic integers. Even more striking for the modern reader is that Hilbert does not employ general abstract notions from group theory that could have unified the discussions of various situations which we immediately recognize as analogous. For instance, he did not heed Minkowski's advice, given in a letter of 21 July 1896 [Minkowski, 1973, 83] to group together at the beginning of art. 100 all lemmata about finite Abelian groups needed in the proof of the so-called Kronecker–Weber theorem (Satz 131).

Similarly, no formal notion of quotient group is used in the *Zahlbericht*, even though the concept of factor group had been first defined and used by Otto Hölder as early as 1889 and discussed in the second volume of Heinrich Weber's *Lehrbuch der Algebra* of 1896 (§53). Thus, when we would say that 'G/H is cyclic of order h', Hilbert writes elaborate prose such as 'The members of G are each obtained precisely once when we multiply the members of H by $1, g, \ldots, g^{h-1}$ where g is a suitably chosen member of G'; see, for example, Sätze 69, 71 and 75. It is remarkable to note by comparison that the 33-year-old Kronecker, while generalizing Gaussian periods to roots of unity of composite order, encountered subgroups H of $(\mathbf{Z}/m\mathbf{Z})^*$ such that the quotient $(\mathbf{Z}/m\mathbf{Z})^*/H$ is cyclic, and added that this property is 'at the same time so characteristic that it could be used as the definition' of such subgroups ([Kronecker, 1856, 33f]; I thank B. Petri for pointing this out to me).

4 CONTENT AND STRUCTURE

The contents of the *Zahlbericht* are summarised in Table 1. We have already made a few comments on its first Part, which contains the basic arithmetic theory of a general finite extension of the field of rational numbers: integers, ideals, discriminant, units, ideal classes, the relationship of the class number with the residue at $s = 1$ of the zeta-function of the field, *Zahlringe*, that is, orders.

The second Part deals with the decomposition of primes in a Galois extension: decomposition group and inertia group, and the corresponding subfields. This theory had been essentially developed but not published by Dedekind, and later independently worked out and published by Hilbert in 1894. Georg Frobenius and Dedekind in their correspondence of February 1895 vented their anger about the fact that Hilbert had failed to acknowledge Dedekind's priority, even though Dedekind had sent Hilbert an offprint in June 1892 explicitly indicating his unpublished work. But Dedekind never published a complaint about Hilbert like the one he wrote against Hurwitz [Dedekind, 1895]. The exposition of this theory in the *Zahlbericht* (arts. 36–47) follows Hilbert's 1894 paper to a large extent.

Table 1. Summary of Hilbert’s report. 372 pages.

Chs.	**Arts.**	**Thms.**	**Topics**
Preface.			Modern number theory and its role in mathematics.
			Part 1: *Theory of a general number field*.
1	1–3	1–5	Number fields; algebraic integers; norm, different, discriminant, integral basis.
2	4–6	6–16	Ideals; decomposition into prime ideals; forms (in Kronecker’s sense).
3–4	7–13	17–37	Congruences mod. an ideal; the discriminant and its divisors; the fundamental equation and unit form of a number field.
5	14–16	38–41	Relative extensions of number fields.
6	17–21	42–48	Units of a number field.
7	22–29	49–57	Ideal classes; class number and the residue of the zeta-function at $s = 1$; characters of an ideal class.
8–9	30–35	58–66	Classes of forms; orders; modules.
			Part 2: *The Galois number field*.
10–11	36–47	67–80	Decomposition in the presence of the Galois group acting; decomposition and inertia groups and fields; powers of the prime divisors of different and discriminant.
12–13	48–52	81–87	Subfields, densities of primes, and composita.
14	53	89	Class group generated by primes of degree one.
15	54–58	90–94	Relative cyclic extensions.
			Part 3: *The quadratic number field.*
16–20	59–90	95–116	Quadratic and norm residue symbol; genus theory; analytic class number formula. Class fields and complex multiplication *not* treated in the report.
			Part 4: *The cyclotomic field.*
21–22	91–98	117–127	Degree, integral basis, discriminant, decomposition, units and circular units.
23	99–104	128–131	All Abelian number fields are cyclotomic.
24	105–112	132–138	Normal bases and root numbers (‘Gauss sums’).
25	113–115	139–140	Eisenstein’s reciprocity law for lth powers.
26–27	116–124	141–146	Cyclotomic analytic class number formula; cyclotomic theory applied to quadratic fields.

Table 1. (*Continued*)

			Part 5: *Kummer's number field.*
28–30	125–135	147–152	Power and norm residue, local symbols, logarithmic derivatives of units; prime ideals with prescribed characters.
31–34	136–165	153–167	The regular Kummer field; ideal classes, genus theory; lth power reciprocity; product formula for norm residue symbols.
35	166–171		Re-arrangement of the preceding theory of the regular Kummer field, avoiding logarithmic derivatives.
36	172–173	168–169	Fermat's Last Theorem for regular prime exponents.

At the end of the second Part, one finds a series of theorems first stated and proved in this generality in the *Zahlbericht*, and which are remarkable for their later impact: Satz 89–94. In Satz 89, Hilbert gives a *non-analytic* proof for the fact that the ideal class group is *generated* by the classes of prime ideals of degree 1. This theorem and its proof have apparently not received the attention they deserve; it took 80 years to see that the proof had to be completed in a technical point [Washington, 1989].

Hilbert's Satz 90, itself a literal generalization of a slightly more special result and proof of Kummer's, has become a household name since the introduction of Galois cohomology in the 1950s. This reinterpretation—which transforms Hilbert's explicit statement into the triviality of a first cohomology group: $H^1(G, K^*) = 1$—along with the substantial generalisation from cyclic to abelian extensions K/k (and many even more substantial generalizations or analogues in later developments), was first initiated by Noether in her work on what was then called the 'Principal Genus Theorem' [Noether, 1933]. To be sure, she translated into the calculus of cross product algebras; the further translation into Galois cohomology came later [Lemmermeyer, to appear].

Satz 91 on the existence of relative units was to be the first in a series of generalizations of Dirichlet's Unit Theorem; and Sätze 92–94 have been forerunners of important results in class field theory. For slightly more detailed comments on these and other mathematical points, see the introduction to the English translation of the *Zahlbericht* by Lemmermeyer and Schappacher.

The third Part of the *Zahlbericht* deals with quadratic fields. Gauss's genus theory (Satz 100) is treated via the Hilbert symbol, that is, the local norm residue symbol that is the main systematic novelty that Hilbert introduced into the treatment of algebraic number theory: he shifted the emphasis from the question, whether a given element is an lth power, to the question of whether it is the norm of an element in a certain extension of degree l. Quadratic reciprocity (Satz 101) is also couched in terms of Hilbert's symbol. This Part also contains the analytic class number formula in the quadratic case, as well as a discussion of arbitrary orders in quadratic fields and their relation to quadratic forms.

The theory of cyclotomic fields follows suit in the fourth Part, including the theory of circular units, and together with Hilbert's proof of the so-called Kronecker–Weber Theorem (Satz 131) to the effect that every abelian extension of the rational numbers is con-

tained in a suitable cyclotomic field. Hilbert had actually been the first mathematician to have published (in 1896) a complete proof of this conjecture by Kronecker [Neumann, 1981, 125].

Then follows a largely original discussion of normal bases and what Hilbert calls their ‘associated root numbers’, that is, generalized Gaussian sums. The prime decomposition of Gaussian sums was obtained in fair generality by Ludwig Stickelberger. Hilbert quotes this article, but only in the context of quadratic fields and not in this section where he derives his own results towards the decomposition of root numbers (Satz 133, 134) and never gives more than a special case of Stickelberger’s theorem (Satz 138), which was already known to C.G.J. Jacobi and Kummer. Helmut Hasse’s incidental complaint about ‘Hilbert’s inconceivably not giving [Stickelberger’s result] in his Zahlbericht’ (letter to Harold Davenport, 22 February 1934) indicates how much later number theorists relied on Hilbert’s report as a comprehensive reference for the 19th-century literature. The subject of root numbers has developed into an active field of research only in the last 30 to 40 years.

The *Zahlbericht* culminates in the long fifth and last Part on *the Kummer number field*. Hilbert describes it in the preface as

> the theory of those fields which Kummer took as a basis for his researches into higher reciprocity laws and which on this account I have named after him. It is clear that the theory of these Kummer fields is the highest peak reached on the mountain of today’s knowledge of arithmetic; from it we look out on the wide panorama of the whole explored domain since almost all essential ideas and concepts [. . .] find an application in the proof of the higher reciprocity laws.

Concretely, the Kummer field is obtained by adjoining to the rational number field all lth roots of unity and an lth root of an element of this cyclotomic field which is not an lth power. The theory works all the way for regular prime numbers l. It is especially in this Part that Hilbert’s struggles with Kummer’s formidable ‘computational apparatus’. In fact, he does the whole theory twice over: the first time around (essentially arts. 131–165), he defines the local norm residue symbol directly and uses Kummer’s device of logarithmic derivatives of circular units to derive its relevant properties at the bad places. The major stepping stone on the way to the general reciprocity law is Eisenstein’s reciprocity law which relates a rational to an arbitrary cyclotomic integer. Although this presentation already reduces ‘Kummer’s computational devices to a small amount’ (art. 166), Hilbert then does go on to rearrange the theory ‘in a way, completely avoiding those computations’ (arts. 166–171). The trick is to use the product formula to recuperate the information needed at the bad places from those at the good ones. Either way, the reciprocity laws are developed along with genus theory for the Kummer fields, and Hilbert treats genus theory via ‘characters’ defined in terms of suitable local norm residue symbols. This feature as well as several technical improvements account for the difference, and in fact superiority of Hilbert’s presentation over Kummer’s genus theory.

The *Zahlbericht* ends with a proof of Fermat’s ‘last theorem’ (in a generalized form) for regular prime exponents (art. 172), and other special cases of it (art. 173).

5 LATER REACTIONS

Later commentators have reacted differently to Hilbert's *Zahlbericht* in general and to his treatment of Kummer's achievements in particular. Major number theorists of the following generation like Erich Hecke and Hasse either learned their number theory from the *Zahlbericht* or used it as a standard reference. Even mathematicians like Felix Hausdorff and Hermann Weyl, whose principal research interests were far from number theory, were influenced by it. Hausdorff for instance, in his letter of congratulations to Hilbert's 70th birthday, wrote: 'My preferred dish among all the delicate things you have served us is the Zahlbericht. It is the most lucky blend of past, present, and future (the three dimensions of time, according to Hegel): the perfect command and exposition of the past, the solution of new problems, and the most refined prescience of things to come'.[1]

In his 1922 praise of 'The algebraist Hilbert', Otto Toeplitz (himself not a number theorist) went as far as writing that 'Hilbert has extracted from Kummer's difficult works overflowing with inductive material, which few before him had read, and which only few will now have to read after him and thanks to him, a universe of general facts and theses' [Toeplitz, 1922, 73]. Hasse in 1932 (in Hilbert, *Gesammelte Abhandlungen*, vol. 1, 529) and Emil Artin in 1962 [Artin, 1965, 549] acknowledged, more soberly than Toeplitz, the conceptual simplification and clarification of Kummer's theory obtained by Hilbert. On the other hand, in section 3 above we have mentioned and tried to interpret Noether's criticism of the *Zahlbericht* from the 1930s.

In 1975, André Weil wrote [Kummer, 1975, 1]:

> The great number-theorists of the last century are a small and select group of men. ... Most of them were no sooner dead than the publication of their collected papers was undertaken and in due course brought to completion. To this there were two notable exceptions: Kummer and Eisenstein. Did one die too young and the other live too long? Were there other reasons for this neglect, more personal and idiosyncratic perhaps than scientific? Hilbert dominated German mathematics for many years after Kummer's death. More than half of his famous *Zahlbericht* [...] is little more than an account of Kummer's number-theoretical work, with inessential improvements; but his lack of sympathy for his predecessor's mathematical style, and more specifically for his brilliant use of p-adic analysis, shows clearly through many of the somewhat grudging references to Kummer in that volume.

Even though the polemical evaluation of Hilbert's toiling as 'inessential improvements' clearly reflects Weil's personality, as does the intentional anachronism to speak of 'p-adic methods' in the middle of the 19th century, his opinion is surely best understood in the context of the renaissance of Kummer's ideas and techniques in the wake of the development of Iwasawa theory, which started in the 1960s and continues to this very day. But all

[1] 'Meine Lieblingsspeise unter all den Delikatessen, mit denen Sie uns bewirtet haben, ist der Zahlbericht. Das ist die glücklichste Mischung zwischen Vergangenheit, Gegenwart und Zukunft (den drei Dimensionen der Zeit, nach Hegel): vollendete Beherrschung und Darstellung des bereits Geleisteten, Lösung neuer Probleme, und feinstes Vorgefühl für die kommenden Dinge' (Göttingen University Library Archives, Cod. Ms. Hilbert 452c, Nr. 15, 21 January 1932). I thank Walter Purkert for having communicated this letter.

these fairly recent developments did of course occur on the firm basis of a well-established algebraic number theory (and class field theory), to the consolidation of which no other single publication has contributed more than Hilbert’s *Zahlbericht*.

BIBLIOGRAPHY

Artin, E. 1965. *Collected papers*, Reading, MA: Addison Wesley.

Blumenthal, O. 1935. ‘Lebensgeschichte’, in Hilbert, *Gesammelte Abhandlungen*, vol. 3, Berlin: Springer, 388–429.

Boniface, J. and Schappacher, N. 2001. ‘ “Sur le concept de nombre en mathématique”—cours inédit de Leopold Kronecker à Berlin (1891)’, *Revue d’histoire des mathématiques, 7*, 207–275.

Brewer, J.W. and Smith, M.K. (eds.) 1981. *Emmy Noether, a tribute to her life and work*, Basel: Dekker.

Dedekind, R. 1895. ‘Über die Begründung der Idealtheorie’, *Nachrichten der Königlichen Gesellschaft der Wissenschaften zu Göttingen*, 106–113. [Repr. in *Gesammelte mathematische Werke*, vol. 2, Braunschweig: Vieweg, 1932, 50–58: cited here.]

Kronecker, L. 1856. ‘Über die algebraisch auflösbaren Gleichungen’, *Monatsberichte der Berliner Akademie der Wissenschaften*, 365–374. [Repr. in *Werke*, vol. 4, Leipzig: Teubner, 1929, 27–37.]

Kronecker, L. 1882. ‘Grundzüge einer arithmetischen Theorie der algebraischen Grössen’, *Journal für die reine und angewandte Mathematik, 92*, 1–122. [Repr. in *Werke*, vol. 2, Leipzig: Teubner, 1895, 239–387.]

Kummer, E. 1975. *Collected works*, vol. 1 (ed. A. Weil), Berlin: Springer.

Lemmermeyer, F. To appear. ‘The development of the principal genus theorem’, in *The shaping of arithmetic after C.F. Gauss’s Disquisitiones Arithmeticae*, Berlin: Springer.

Minkowski, H. 1973. *Briefe an David Hilbert* (ed. L. Rüdenberg and H. Zassenhaus), Berlin: Springer.

Neumann, O. 1981. ‘Two proofs of the Kronecker–Weber theorem “according to Kronecker, and Weber”’, *Journal für die reine und angewandte Mathematik, 323*, 105–126.

Noether, E. 1933. ‘Hauptgeschlechtssatz für relativ-galoissche Zahlkörper’, *Mathematische Annalen, 108*, 411–419.

Reid, L.W. 1910. *The elements of the theory of numbers*, New York: MacMillan. [With an introduction by David Hilbert.]

Sommer, J. 1907. *Vorlesungen über Zahlentheorie, Einführung in die Theorie der algebraischen Zahlkörper*, Leipzig und Berlin: Teubner.

Toepell, M.-M. 1986. *Über die Entstehung von David Hilberts “Grundlagen der Geometrie”*, Göttingen: Vandenhoeck und Ruprecht.

Toeplitz, O. 1922. ‘Der Algebraiker Hilbert’, *Die Naturwissenschaften, 10*, no. 4, 73–77.

Washington, L.C. 1989. ‘Stickelberger’s theorem for cyclotomic fields, in the spirit of Kummer and Thaine’, *Théorie des nombres 1987*, Quebec, Canada, 990–993.

CHAPTER 55

DAVID HILBERT, *GRUNDLAGEN DER GEOMETRIE*, FIRST EDITION (1899)

Michael Toepell

In mathematics it was this influential work that led by its axiomatic method to a new thinking in all mathematical fields in the 20th century. In addition, following in the traces of Euclid, it became the classical textbook for geometry in educating mathematicians and mathematics teachers for nearly the whole century.

First publication. Part 1 of *Festschrift zur Feier der Enthüllung des Gauß–Weber-Denkmals in Göttingen*, Leipzig: Teubner, 1899. 92 pages.

Manuscripts. Main manuscripts are preserved in the *Niedersächsische Staats- und Universitätsbibliothek*, Göttingen *Cod. Ms. D. Hilbert* (see Table 2 below).

Later editions. 2nd 1903, 3rd 1909, 4th 1913, 5th 1922, 6th 1923, 7th 1930, 8th 1956, 9th 1962, 10th 1968, 11th 1972, 12th 1977, 13th 1987, 14th ed. (ed. M. Toepell), 1999. All Teubner. [3rd-7th eds. in the series 'Wissenschaft und Hypothese', vol. 7. 8th, 9th and 11th eds. ed. P. Bernays. 10th, 12th and 13th eds. unchanged from the preceding eds. 14th ed. with commentaries and extensive bibliography.]

French translation. 'Les principes fondamentaux de la géométrie' (trans. L. Laugel), *Annales scientifiques de l'Ecole Normale Supérieure, (3) 17* (1900), 103–209.

English translation. The foundations of geometry (trans. E.J. Townsend), Chicago: Open Court, 1902. [Repr. 1947. 2nd ed. (of the 10th ed.; trans. L. Unger) La Salle, Ill.: Open Court, 1971.]

Related articles: Grassmann (§32), Riemann on geometry (§39), von Staudt (§33), Klein (§42), Hertz (§52), Hilbert on algebra (§54), Einstein (§63), Hilbert and Bernays (§77).

1 A NEW DIRECTION OF MATHEMATICAL THOUGHT IN 1899

Well into his scientific career, David Hilbert (1862–1943) published most his decisive and influential work in 1899. It was surprising to his contemporaries that it did not deal with

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

his common topics, algebra or number theory, but with geometry. The work impressed methodically by its conscious lack of intuition, of perception of space (*Anschauung*) and geometric experiments. This was absolutely unusual at that time.

Hilbert contributed to a better understanding of the coherence of geometric and algebraic structures, continuing some ideas laid out in Felix Klein's 'Erlanger Programm' 1872 (§42). With this work the foundations of geometry was established as a field of research in itself.

David Hilbert, born on 23 January 1862 in Königsberg i.Pr. (nowadays Kaliningrad in Russia) to a country judge, studied mathematics and physics at the universities of Königsberg, Heidelberg, Leipzig and Paris. He took his doctoral degree in 1884 at Königsberg with Ferdinand Lindemann (1852–1939), who was well known for his proof of the transcendence of π. After his *Staatsexamen* in 1885 and his *Habilitation* in 1886 Hilbert became *Privatdozent*, in 1892 *ausserordentlicher Professor* and from 1893 *ordentlicher Professor* at the University of Königsberg. In 1895 Klein managed to obtain Hilbert for Göttingen University, where he lived until his death on 14 February 1943.

Beginning with his dissertation about algebraic invariants, up to 1899 Hilbert was well-known in mathematics as an expert on algebraic theory of invariants and on number theory (§54). From 1904 he became concerned with integral functions, mathematical physics and logic, he discovered new fundamental results in these fields. In the meantime he had an intense and profound geometric period. Let us trace his thoughts in this respect. What happened at that time of the beginning of modern mathematics?

Towards the end of the 19th century a remarkable change came about in the field of the foundations of geometry. Whereas geometry had hitherto been based on empirical facts, it was now seen as a purely formal deductive system. Hilbert was not the first so to act, but he perfected this method in his book *Grundlagen der Geometrie*. It also helped to lead him to study mathematical logic in the early 1900s (compare §61).

2 FOURTEEN EDITIONS IN 100 YEARS

The contents of Hilbert's book are summarised in Table 1. It first appeared in June 1899 in a *Festschrift* commemorating the unveiling of the Gauss–Weber-Monument in Göttingen; he did not give a lecture on that occasion. The 5th edition appeared in 1922 in the year of his 60th birthday, the 9th edition in 1962 for his centenary, and the 12th edition in 1977 for the bicentenary of C.F. Gauss's birth.

The centenary of the book in 1999 was an occasion to revise and complete the book. So a *Jubiläumsausgabe* appeared, containing contributions that summarize the pre-history and the further development in the last hundred years, as well as documents and registers, which complete and organize the book. It contains a remarkable exercise book on the foundations of projective geometry written when Hilbert was still a school-boy, an appreciation of the omitted projective geometry, a critical comparison of all the 13 editions, a register of literature and names, numerous photographs, title-pages and facsimiles. For a mathematical university-level textbook in modern times it is quite unusual in appearing over 100 years in 14 editions.

Table 1. Summary by Sections of *Grundlagen der Geometrie* (1899).

Sect.; pp.	**'Title': other included topics**
I; 4–19	'The five groups of axioms': connection, of order, of parallels, of congruence, of continuity; consequences.
II; 19–26	'The compatibility and the mutual independence of the axioms': compatibility, independence, non-Euclidean and non-Archimedian geometry.
III; 26–39	'The theory of proportion': complex number-systems, arithmetic of segments.
IV; 40–49	'The theory of plane areas': equal areas, equal content, measure of area.
V; 49–71	'Desargues's theorem': provability, new segment arithmetic, algebraic laws.
VI; 71–77	'Pascal's theorem': provability, non-Archimedean number sets.
VII; 78–88	'Geometrical constructions based upon the Axioms I–V': geometric constructions with a ruler and a transferer of segments.
–; 89–92	'Closing word'; table of contents.

3 SEEING THE MASTER IN HIS WORKSHOP: HILBERT'S MANUSCRIPTS

Hilbert's method immediately gave a new direction to mathematical thought in the 20th century. Its impact on contemporaries has been studied and further developed in numerous publications (A. Schmidt in [Hilbert, *Papers*, vol. 2, 404–414]; [Freudenthal, 1957]; and B.L. van der Waerden in the foreword of the later editions of the *Grundlagen*). However, up to the 1980s little was known about the origins of the book, or about the developments that led him to it. Written in Hilbert's typical concise manner, the book itself offers nearly no information on this subject.

According to the biographies [Blumenthal, 1935; Reid, 1970; Dehn, 1922; Weyl, 1944], Hilbert appeared to have worked almost exclusively with algebra and questions concerning the theory of numbers in the years prior to 1899. His publications on the theory of invariants suggest as much. With the *Grundlagen der Geometrie*, however, he presented to the public a thoroughly mature work about an entirely different subject. In 1935 Otto Blumenthal (1876–1944), Hilbert's first assistant in Göttingen, wrote in his biographical sketch of Hilbert that the book 'has brought up to its author a world-wide reputation, whereas up to that time he was appreciated only among experts. It is worth tracing the grounds for this success and the development of Hilbert's ideas' [Blumenthal, 1935, 402].

How did Hilbert arrive at his creation? What kind of preparatory work did he find? What works did he study? Which problems had particularly stimulated him? These matters were not investigated during the following 50 years. One reason is that his manuscripts were not made accessible until 1973, 30 years after his death. The catalogue covers over 700 items, among them letters from roughly 500 correspondents and about 50 manuscripts of his own lectures. Now it was possible not only to describe the mere content of the book but also to follow Hilbert's way of constructing it [Toepell, 1986a]. From our viewpoint the principal

Table 2. Pertinent manuscripts

Abbr.	**Manuscripts of Hilbert's lectures**	**Semester**	**Cod.**
PG	Projektive Geometrie.	Summer 1891	535
GG	Die Grundlagen der Geometrie.	Summer 1894	541
FK	Ueber den Begriff des Unendlichen (Ferienkurs).	Easter 1898	597
EG	Grundlagen der Euklidischen Geometrie.	Winter 1898–99	551
SG	Elemente der Euklidischen Geometrie (Ausarbeitung von Hans von Schaper).	March 1899	552

correspondents are Felix Klein, Hermann Minkowski (1864–1909), Ferdinand Lindemann, Adolf Hurwitz (1859–1919) and in later times Albert Einstein (1879–1955).

A first complete set of axioms before Hilbert was constructed in [Pasch, 1882], who also set up the axioms of order that C.F. Gauss had already postulated [Contro, 1976]. Manuscripts of four Hilbert lecture courses form the basis of his development (Table 2).

Hilbert's first lectures on geometry dealt with *Projektive Geometrie* (1891, hereafter 'PG'). They dealt with the properties that are invariant under projections. Then came a manuscript on non-Euclidean geometry, axiomatically formulated; *Die Grundlagen der Geometrie* of 1894 ('GG'). The third source relates to an Easter vacation course in 1898, *Ueber den Begriff des Unendlichen* ('FK'); it seems to form the kernel of the later book. From it emerged the detailed manuscript on Euclidean geometry, written in the winter semester 1898–1899: *Grundlagen der Euklidischen Geometrie* ('EG'). Finally, the elaboration *Elemente der Euklidischen Geometrie* ('SG'), was prepared by Hilbert's assistant Hans von Schaper from the preceding lectures in March 1899; and from that text Hilbert developed the book, which was published in June 1899.

These sources made it possible to demonstrate the origins of the book, and also to trace the steps that Hilbert had omitted from his publications. These concern, first of all, the role of intuition (perception, *Anschauung*), of experience and experiments, as well as questions of projective geometry, which are omitted from the book. We also see the close connection between Hilbert's biography and the genesis of axiomatic thinking.

4 FOUNDATIONS OF PROJECTIVE GEOMETRY: HILBERT'S EXERCISE-BOOK (1879) AND LATER

Of the time when Hilbert was a schoolboy there are two exercise books of special interest for us. One of them, from 1879 (published in 14th ed., 327–345), deals with projective geometry and the point of intersection theorems, and can be seen as the starting point of his geometric studies. It deals with the theorems of Menelaos and Ceva; cross ratios; arithmetic, geometric und harmonic means; the complete quadrilateral; polar theory; the position of the five remarkable points of a triangle; and the theorems of Pascal and Brianchon.

Unlike the later manuscripts this exercise book is tidy, with the diagrams drawn rather exactly. The sections about the harmonic points and rays and about Pascal's theorem were

to be elaborated in Hilbert's later lectures on projective geometry in summer 1891. That manuscript, PG, comprises over 100 pages. He began with a fundamental survey, dividing geometry into three parts, a division that he consistently followed in later years (fundamental, but not mentioned in the *Festschrift*).

On *Intuitive geometry* Hilbert included school geometry, projective geometry and the *Analysis Situs* (topology); the aim was aesthetic, pedagogical and practical. From 1920 Hilbert delivered a lecture course several times on 'Anschauliche Geometrie', which was published as [Hilbert and Cohn-Vossen, 1932]. On *Axioms of geometry* 'This part investigates, which axioms are used in the established facts in intuitive geometry and confronts these systematically with geometries in which some of these axioms are dropped'; this was epistemological ('*erkenntnistheoretisch*') in intent. These considerations led him to investigate independence and to his own geometries in the manuscript GG of 1894. In *Analytical geometry*, 'from the outset a number is ascribed to the points in a line and thus reduces geometry to analysis'; this was 'scientifically mathematical' ('*wissenschaftlich mathematisch*').

In 1891 Hilbert still based 'geometry' on experiments and ontological facts. He begins his lecture: 'Geometry is the theory about the properties of space. [...] It is based on the simplest experiment that can be carried out, namely drawing' [Toepell, 1986a, 21]. As a consequence Hilbert excluded computing and numbers.

Hilbert mentioned the *Geometrie der Lage* (1847) of K. von Staudt (§33), following its pure methods in order to keep projective geometry free from axiomatic and analytic influences. In the first part Hilbert followed Theodor Reye's *Geometrie der Lage* (1866, 3rd edition 1886), and in the second part Jakob Steiner's lectures on synthetic geometry as edited by H. Schroeter (1866, 3rd edition 1898).

At the end of September 1891, Hilbert heard a lecture on geometry given by Hermann Wiener at the annual congress of natural scientists in Halle [Wiener, 1891]. Thereby Hilbert became acquainted with the general validity of the axiomatic method and in particular with the possibility of developing projective geometry by taking as axioms Pascal's and Desargues's theorems on point of intersection. According to [Blumenthal, 1935, 402], Hilbert uttered these famous words in a Berlin waiting-room on the return journey from Halle to Königsberg: 'One should always be able to say, instead of "points, lines and planes", "tables, chairs and beer mugs"'. If this report is reliable, already in 1891 he saw the intuitive part of geometrical concepts as being mathematically irrelevant. But only seven years later he did express that view in a correspondingly radical formulation.

5 GEOMETRY AS A SYSTEM OF AXIOMS (1894)

A key point for Hilbert was the construction of a system of *independent* axioms. After the study of the role of the axiom of parallels with Gauss, J. Bolyai and N.I. Lobachevsky, it was at least Hermann Grassmann who demanded in 1844 not to have any unnecessary axioms (§34). Giuseppe Peano was the first to speak of the concept of *independence* of axioms [Peano, 1889, 57].

In the summer semester of 1894 Hilbert gave his lectures on non-Euclidean geometry under the title *Die Grundlagen der Geometrie* ('GG'). He wanted to produce the purest

possible exact system of axiomatic, non-Euclidean geometry, concluding with Euclidean geometry. He prefaced his manuscript with a bibliography of over 40 items (most unusual in Hilbert's work) available in German. Amongst others he named Pasch, H. von Helmholtz, Lobachevsky, Bernhard Riemann, Peano, W. Killing, Sophus Lie, R. Clebsch, Lindemann, A.F. Möbius, von Staudt, Reye, B. Erdmann and Wiener. He mentioned Italian works as far as they were translated, such as [Peano, 1891]. But he did not mention the important axiomatic studies [Peano, 1889] or [Fano, 1892]. Peano's *Sui fondamenti della geometria* (1894) had only just come out, as also Giuseppe Veronese's *Fondamenti di geometria a più dimensioni* (1891) in German, as *Grundzüge der Geometrie von mehreren Dimensionen* (1894) just been translated. In 1899 Hilbert mentioned non-Archimedian geometry that Veronese tried to construct; in the first to the sixth editions of his *Festschrift* he noted Veronese's remarkable historical appendix.

The manuscript of 1894 was a further step to the *Festschrift*. Contrary to 1891, Hilbert avoided any explicit definition of geometry like 'Geometry is the theory about the properties of space'. Otherwise physical properties, like the falling principles, would belong to geometry, too. So he wrote: 'Among the phenomena, or facts of experience that we take into account observing nature, there is a particular group, namely the group of those facts which determine the *external form of things*. Geometry concerns itself with these facts' [Toepell, 1986a, 58]. He even regarded the axioms as *facts*: 'These unprovable facts have to be determined in advance and we term them axioms'.

Here Hilbert still stood at the same level with Pasch, who likewise derived his axioms from 'experience'. However, Hilbert also questioned whether the axioms are *complete* or *independent*: 'Our colleague's problem is this: which are the necessary and sufficient *conditions*, independent of each other, which one must posit for a system of things, so that every property of these things corresponds to a geometrical fact and vice versa, so that by means of such a system of things a complete description and ordering of all geometrical facts is possible'. In addition, he took up an idea from Heinrich Hertz's *Die Prinzipien der Mechanik* (1894), which in geometry leads to the use of space intuition ('*Raumanschauung*') only in the sense of a possible intuitive *analogy* (§52.6).

Whereas Pasch [1882] divided his axioms into eight axioms for lines and four for planes, Hilbert now arranged his axioms according to *relations*, an order he had already touched on in 1891. He first separates the axioms of connection and order. Following the practice of von Staudt and Möbius, he now ascribed rational numbers to the point on a line with the help of the construction of the fourth harmonic element. In order 'to prevent a gap' in the transition to the real numbers, he stated an axiom of *continuity*; from Pasch he adopted the formulation given by Karl Weierstrass.

Subsequently the fundamental theorems of projective geometry were derived. Then Hilbert focussed upon the projective system. This led to the axioms of congruence, and the determination of metrics to hyperbolic and parabolic geometries. The second part was based on the non-axiomatically constructed *Vorlesungen über Geometrie* of Clebsch as edited by Lindemann in 1891. A main problem for Hilbert was to select and to formulate the suitable axioms.

In retrospect it is remarkable that Hilbert states the axiom of *continuity* straight after the first two groups of axioms. Together with the *introduction of numbers* this, for him, was 'of high epistemological significance'. He had not anticipated from the outset the early

introduction of numbers. In the second part of the lectures he often proceeded analytically, dividing geometry in three different parts in a way that was logical but for his axiomatic system impure, which he wished to avoid in future. In this regard he noted: 'If I lecture again, it will be on Euclidean geometry' [Toepell, 1986a, 104]; and so he did, in 1898. Thus it can be understood why he added the axioms of continuity at the *end* and thus showed how dispensable they were.

How could Hilbert get algebraic laws by geometrical means, or by axioms? This is possible by the power of the intersection theorems of Desargues and Pascal. They enabled him to establish a *segment arithmetic* (that is, an arithmetic of geometrical entities) without an axiom of continuity. In 1894 he had not even mentioned the intersection theorems and consequently had not examined their importance. Thus [Wiener, 1891], which had been seen as decisive by Blumenthal, was not applied in this way by Hilbert before 1898!

In addition, Hilbert viewed continuity as one of the assumptions of *projective geometry*. As he tried to avoid continuity, the projective studies disappeared from his *Grundlagen der Geometrie*. Other reasons for the transition from projective to Euclidean geometry were that the order relation of three elements had proved unsuitable in projective geometry, and that the principle of duality is not valid in the geometry that he outlined in 1894. Due to Hilbert's book of 1899, in the following decades projective geometry also gradually disappeared in school geometry.

6 A VACATION-COURSE FOR TEACHERS: THE KERNEL OF THE *FESTSCHRIFT* (1898)

At Easter in 1895 Hilbert accepted the chair at Göttingen and up to 1897 he concerned himself principally with number theory. So his concern with the foundations of geometry rested for more than three years, until he was inspired to take it up anew by a letter of 30 January 1898 sent by Friedrich Schur to Klein. Hilbert wrote in March 1898 to Hurwitz: 'This letter, which [Artur] Schoenflies introduced to us in a lecture to the mathematical society, has given me the inspiration to take up again my old ideas about the foundations of Euclidean geometry. It is remarkable how many new things can be discovered in this field' [Toepell, 1985, 641].

The manuscript (FK) of an Easter vacation course of 1898, *Ueber den Begriff des Unendlichen* ('On the concept of infinity'), covers just 27 pages; but it forms the nucleus of the *Festschrift*. As we see from the introduction to this course, the contact with teachers and school-mathematics was Hilbert's personal request. He addressed himself especially to the teachers as 'the most competent collaborators'; perhaps they who especially stimulated him to study the foundations of geometry. A reviewer even postulated that the *Grundlagen der Geometrie* should be used as a textbook in school geometry, like a new Euclid. Indeed, it became a fundamental textbook in university geometry in the 20th century.

In this course Hilbert introduced his audience to the most up-to-date research questions. For the first time he constructed the axioms in what was subsequently to be their usual sequence. Then he directed the teachers to practical problems: the geometrical constructions based on the theorems of congruence. For example, the constructibility of the intersecting point of two circles required an axiom of continuity, whose independence was subsequently

examined. Also he asked for the first time, which axioms were dispensable if one assumed Desargues's and Pascal's theorems in place of some axioms that were used to prove these theorems.

In this manuscript the arrangement of the later *Festschrift* is already apparent: axioms, proofs of independences, segment arithmetic, Desargues's theorem, Pascal's theorem, and problems concerning constructibility. We can also trace how Hilbert developed his ideas in two directions: to avoid assumptions of *continuity*, and to construct plane geometry independent of *spatial* assumptions. Once Hilbert's basic concept had been established, a number of individual problems came into focus on which he now worked intensively. That led him to the careful system in the *Festschrift*.

7 LECTURES AND AN ELABORATION ON EUCLIDEAN GEOMETRY (1898–1899)

In the winter semester of 1898–1899 we read in the announcements of lectures in Goettingen: 'Elemente der Euklidischen Geometrie: Prof. Hilbert, Montag und Dienstag 8–9 Uhr, privatim'. So, two hours per week. Hilbert began: 'Concerning the content of the lectures, we shall study the theorems of elementary geometry, which we all learned at school: the theory of parallels, the theorems of congruence, the equality of polygons, the theorems about the circle etc. in the plane and the space' [Toepell, 1986a, 144].

The manuscript EG contains an exhaustive discussion of those areas that were mostly treated in brief in the vacation course. The *logical meaning* of the axioms was studied by construction of arithmetical models. Amongst these were proofs of independence for axioms of the first two groups. In accordance to the theme of the lectures, Hilbert examined in detail the studies of congruence that were possible without using continuity. Much of this was omitted in the *Festschrift*, including (unusually in his writings) an *historical* survey of the parallel axiom that follows, then the detailed presentation of a *non-Euclidean* geometry and the introduction of *ideal* (infinite) *elements*.

Comparing this lecture with that of 1894, it is plausible when Klein remarked of the *Festschrift* that 'compared with earlier studies its main object is to state the importance of the axioms of continuity' [Klein, 1914, 402]. Freudenthal [1957] asserted that 'The so-called axioms of continuity are introduced by Hilbert to show that actually they are dispensable'. Because of this important result, Hilbert introduced them at the end.

In March 1899 Hilbert's assistant von Schaper had elaborated these lectures as the text *Elemente der Euklidischen Geometrie* (SG). It contains numerous remarks, motivations and examples, which were omitted in the concise presentation of the *Festschrift*.

Here Hilbert began with the fundamental concepts. He did not explain it as in the lectures 'es giebt ein System von Dingen, die wir Punkte nennen' ('there is a system of things, that we call points'), but formulated it with abstract rigour: 'Zum Aufbau der Geometrie denken wir uns drei Systeme von Dingen, die wir Punkte, Geraden und Ebenen nennen'. In the *Festschrift* he omitted even the words 'zum Aufbau' and 'uns' and wrote: 'Wir denken drei verschiedene System von Dingen'; Leo Unger translated this as 'Consider three distinct sets of objects' in the second English edition of 1971.

'With these lion-claws the navel-string between reality and geometry is cut through' [Freudenthal, 1957, 111]. Geometry seems to awake to its own existence, independent of

any physical reality. Some months before Hilbert had still seen the axioms as 'very simple [. . .] original *facts*' (SG), whose validity is experimentally provable in nature.

What Hilbert formulated may have been new in Germany, but it was "in the air". Already seven years earlier Fano wrote: 'At the basis of our study we put some variety of entities of some nature; entities that we shall call, for brevity, points, but independently, well agreed, of their actual nature' ([Fano, 1892, 108f]; see [Toepell, 1999b, 295]). In the two years following (1892–1894) Fano had been in Göttingen.

Concerning the further development it is interesting that the elaboration included studies of some theorems of Legendre that Hilbert published only thirty years later in the seventh edition (art. 10, 39–45; see [Toepell, 1986a, 208]). An eight-page introduction to projective geometry by means of ideal elements was also omitted by Hilbert from the *Festschrift* [Toepell, 1986a, 212–215].

8 THE ALGEBRAISATION OF GEOMETRY: THE FIRST EDITION IN JUNE 1899

In spring 1899 Hilbert once more revised his lectures, for the *Festschrift*. Now he concentrated his wide-spread investigations upon questions of independence and especially Desargues's and Pascal's theorems in special chapters.

Hilbert's *aim* from the outset seems to be the algebraisation of geometry. In 1894 he still was satisfied with the introduction of coordinates by means of the Möbius grid. At the end he established that it must also be possible to *calculate* with these numbers ascribed to geometrical objects. Hence the algebraic laws of fields ('*Körpergesetze*') were required.

While Pasch had spoken of primitive propositions 'directly based on observation', from which he derived all the remaining theorems, for Hilbert the *relations* between the objects of intuition provide the starting point, as in his manuscript of 1894. Having perceived both the starting-point and the aim, it remains only to find the way. Like Euclid, Hilbert proceeds axiomatically. Here the question arises, which axioms are required?

While the axioms of incidence were largely clear, those of order were somewhat problematical. Hence, because of infinite elements, difficulties attended projective geometry. Proceeding to Euclidean geometry, the concept of congruence could be introduced without hesitation. But then appeared the problem of the intersection theorems, of the axioms of parallels, of the Archimedian axiom, and of continuity—all aspects that are not mutually independent. As the details demonstrated, it was not easy for him to find a suitable way through this maze of axioms.

9 THE FURTHER DEVELOPMENT OF HILBERT'S *GRUNDLAGEN DER GEOMETRIE*

9.1 Continuously revised. The development of Hilbert's *Grundlagen der Geometrie* was not finished with the first edition. The complete content was continuously revised, especially during his lifetime. He considered new results, gave hints in articles and improved his own formulations. The most incisive changes went through the seventh edition (1922), the last edition to appear in his lifetime. The comparison of all editions reveals many differences, elaborated in [Toepell, 1999a].

9.2 Independences. Also by fine and small differences we see Hilbert's effort to choose his words carefully. For example, sometimes in the first edition he spoke of 'Grundthatsachen' ('fundamental facts'), but in later editions of 'Grundsätzen' ('fundamental theorems'). While he wrote in the first edition that 'none of the axioms can be deduced from the remaining ones' (art. 10), later he weakened this to state that 'no essential part of any one of these groups of axioms can be deduced from the others'. As Schmidt remarked, Hilbert preferred conceptual understanding and intuition to logical economy [Hilbert, *Papers*, vol. 2, 407]. An excellent (but not a categorical one) system of geometrical axioms that are *absolutely* independent, was constructed by Oswald Veblen in his dissertation in 1904.

Some modifications of axioms and theorems led to further enlargements. So for example supplement I goes back to a modification of axiom II.4 of Pasch, as suggested by van der Waerden (14th ed., 241f.). A further example is the theorem of four points by E.H. Moore (p. 6), which in the first edition still was an axiom. The proof of theorem 9 (p. 10), which for Hilbert was a proof 'without significant difficulty' (up to the 8th ed., 10) was delivered by G. Feigl only 25 years later (14th ed., 242; [Toepell, 1999a, 297]).

9.3 Axioms of congruence. According to Hilbert the axioms of congruence have been 'the most important and most difficult group' [Toepell, 1986a, 161]; so he had special interest in the functions of these axioms and less in the axiom of parallels. In his review of the *Festschrift* Henri Poincaré concluded: 'Lobachevsky and Riemann rejected the postulate of Euclid, but they preserved the metrical axioms; in the majority of his geometries, Professor Hilbert does the opposite' [Toepell, 1999a, 297]. That means that he tried to reject the metrical axioms. Out of his study of axiom III.5, the so-called *Umklappungssatz* ('reflection theorem'), emerged 'Appendix II'. This axiom plays an important role in the proof of Desargues's theorem in the plane (14th ed., arts. 22–23).

It is remarkable that the whole theory of congruences was shortened from about 20 pages in von Schaper's elaboration to six pages in the *Festschrift*. At the same time Hilbert generalized the title from 'Grundlagen der Euklidischen Geometrie' to 'Grundlagen der Geometrie'.

9.4 Axiom of completeness. The coronation of Hilbert's axioms of continuity is the axiom of completeness. The first time that he embedded it in his *Grundlagen* was in May 1900, in the French edition, after he had called attention to it already in a discourse 'Über den Zahlbegriff' on 12 October 1899. Out of this talk, which appeared in the *Grundlagen* as appendix VI from the third to the seventh edition, emerged supplement I.2 by Bernays in later editions. With this axiom of completeness the non-isomorphic systems became categorical (that is, all systems fulfilling the axioms are isomorphic). This 'axiom about axioms', whose 'logical structure is complicated' (Bernays), is called by Freudenthal [1957] an 'unlucky axiom', but Richard Baldus thought it to be 'the most original achievement by Hilbert in axiomatics'. In the first edition of the English translation of 1902 it reads: 'In other words, the elements of geometry form a system which is not susceptible of extension, if we regard the five groups of axioms as valid' (p. 25).

This axiom of *completeness* was discussed exhaustively and repeatedly changed [Toepell, 1999a, 299f]. The axioms of *continuity* conclude the system. Out of postulating them

at the beginning emerged Hilbert's appendix IV (1902, *Über die Grundlagen der Geometrie*).

9.5 Theories of proportions and areas. What Hilbert did in the third chapter, the theory of proportions, could have been, in the words of [Freudenthal, 1957], 'a Greek ideal: a pure geometric approach' by constructing a coordinate geometry with respect to a field which does not have to be Archimedian—'an original idea with a powerful effect', removing the second stain in Euclid's *Elements*. Earlier contributions are due to von Staudt (1847 and later) and Schur (1891, 1894, 1898). A simplification was delivered by Adolf Kneser (1901, 1904) and further ones were possible by the proofs of Gerhard Hessenberg in 1905 and Johannes Hjelmslev two years later.

The theory of plane areas in chapter IV turns out to be, in Hilbert's words, 'the supposedly most interesting application' of the axioms I to IV and 'one of the most remarkable applications of Pascal's theorem in elementary geometry' because he did not need any axiom of continuity. In chapter V and VI he coordinated the affine plane by the affine form Desargues's theorem. The role of this theorem came into focus.

Not long after the first edition, F.R. Moulton in 1902, T. Vahlen in 1905 and Vahlen and J.H.M. Wedderburn in 1907 constructed further and also simpler non-Desarguesian geometries. Hilbert took up the Moulton example in (7th ed. (1930), 86f.). A systematic treatment of non-Desarguesian geometries started in the early 1930s with the contributions by Ruth Moufang, who developed the algebraic theory of *alternating fields*.

9.6 Theorems of Desargues and Pascal. The so-called *new segment arithmetic* came out of the question how far Desargues's theorem is able to replace the axioms of congruence. This new geometry, in which the commutative law of multiplication does not hold, opened 'the view to a very large, still not investigated area' [Dehn, 1922], which in the following decades led to *non-commutative algebra*. On the further development in the field *Grundlagen der Geometrie* from around the last editions published in Hilbert's lifetime, see (14th ed., 365–384). For Schur in 1901 'the most important result' of the *Festschrift* was chapter VI, in which Hilbert showed that the proof of Pascal's theorem without the use of the axiom of congruence III.5 is only possible with the aid of Archimedes's axiom.

9.7 Elementary geometric constructions. A remarkable station in the history of elementary geometric constructions is the last final VII, the constructions with ruler and scale. The development can be seen in an impressive manner by means of the problem of constructing the missing centre of a circle [Toepell, 1999a, 312–314]. Euclid solved the problem by using a compass and a ruler, Abu al-Wafa (10th century) and Albrecht Dürer in 1525 reduced their tools to a ruler and a compass with a fixed opening. G. Mohr in 1672 and L. Mascheroni in 1797 used a compass alone, while J.H. Lambert in 1774 and Steiner in 1833 required only a ruler and a fixed circle with its centre, and A. Adler deployed parallel- or angle-rulers in 1890. These developments lead straight to Hilbert and his student Michael Feldblum, who showed the possibility of solving the problem with a ruler and a so-called 'transferer of segments', in later editions with ruler and scale, a transferer of a single fixed segment.

10 A SURVEY OF THE INTERSECTION THEOREMS AND THE MOST IMPORTANT RESULTS

The role and function of the intersection theorems are of highest importance for the whole book. Table 3 lists Hilbert's decisive results; the first steps in his manuscripts are set in brackets. Besides this the following results of the historical investigations seem to be the most important.

Firstly, it is a little-known fact that Hilbert studied the foundations of geometry as early as *1891*, and perceived carefully the development in this field. Secondly, a key role is due to Schur, who was responsible for the decisive stimulation of Hilbert in early 1898, which led to an intensive period discussing the foundations of geometry. Thirdly, the significance of intuition (*Anschauung*) for Hilbert was much more important than his publications suggest. Finally, *projective geometry* plays a remarkable role in the pre-history of the *Grundlagen der Geometrie*. We can trace it back until Hilbert's time as a schoolboy. For him projective geometry always belonged to the foundations of geometry, but there are different reasons why he omitted it from the book.

Table 3. Results for intersection theorems.

Abbreviations. L = axioms I and II; L_2 = axioms I and II for the plane; III = axioms of congruence; IV = axiom of existence of parallels (14th ed., 28); IV* = existence and uniqueness of parallels (*ibidem*, 83); V = Archimedian axiom; Des = Theorem of Desargues; Pas = Theorem of Pascal.

Theorem	**Hilbert's stages and results 1898–1899**			
Source	FK	EG	SG	GG, 1st ed.
Desargues (1648): L $\Rightarrow$ Des	FK 19			
Schur (1898): L III $\Rightarrow$ Pas	FK 19		SG 74	
Hilbert: L_2 III IV $\Rightarrow$ Pas	(FK 23)	(EG 92)	SG 108	III art. 14
Hilbert: L_2 III IV $\Rightarrow$ Des	(FK 19)		SG 108	V art. 22
$L_2 \not\Rightarrow$ Des		EG 30	SG 28	
L_2 IV V $\not\Rightarrow$ Des			SG 146	
L_2 III 1–4 IV* V $\not\Rightarrow$ Des				V art. 23
L_2 IV Des $\Rightarrow$ new segment arithmetic		EG 102	(SG 147)	V art. 24
L_2 Des $\Rightarrow$ L		(EG 34)	(SG 32)	V art. 30
L IV* V $\Rightarrow$ Pas			SG 146	VI art. 31
L IV* $\not\Rightarrow$ Pas	FK 26		SG 147	VI art. 31
L IV Pas $\Rightarrow$ every intersection theorem	(FK 27)	EG 104	SG 167	VI art. 31
Hessenberg (1905): L_2 IV* Pas $\Rightarrow$ Des				VI art. 35 (from 3rd ed.)
Hjelmslev (1907): L_2 II $\Rightarrow$ Pas		(EG 106)		III art. 14 (from 3rd ed.)

11 REACTIONS AND CONCLUSION

The *Festschrift* led to the world-wide reputation of Hilbert. The first written congratulations came from Minkowski, Hurwitz and Aurel Voss. One of the first public reactions to the epistemological background can be found in a lecture by Otto Hölder a month after publication, on 22 July 1899. The philosophical discussion about the nature of axioms with Gottlob Frege led to Hilbert's decisive article 'Ueber die Grundlagen der Logik und Arithmetik' in 1905, a philosophical *Programmschrift*, which was reprinted as appendix VII of the book from the third to the seventh editions. Other texts are discussed in [Toepell, 1999a, 316–320].

The book had some consequences for physics, which gained Hilbert's attention from the 1900s onwards. For during the 1910s there was a remarkable connection between Einstein and Hilbert. Seeking the roots and sources of Einstein's ideas in geometry, we are led back to the time of his being a student of Minkowski at the Polytechnical High School in Zurich—especially in 1899, when he had read the proof-sheets of Hilbert's work. According to his letters we may assume that he understood the immense power of this book and it should not take too long for initiating Einstein as well. Maybe Einstein learnt from Hilbert to free himself from empirical restrictions in geometry. This led 12 years later to the idea to think of spatial curvature not only in a Euclidean or non-Euclidean form but also in a form emerging out of gravitational forces (§63).

It is not at all obvious that the conception of the *Grundlagen der Geometrie* emerged from a vacation course for teachers; the significance of intuition seems to be entirely subordinate. Also in his further publications Hilbert argued as a rule for the *axiomatic method*. Hence he was frequently seen as a formalist. However, he never once used 'formalism' to characterise his philosophical position, and his manuscripts and letters show his intense concern with intuition and its significance for geometry. Regarding his attitude in later years we see how little Hilbert freed himself from *intuition*. He perceived that the consistency of his axiomatic system depends after all on what it means.

One hundred years after Hilbert's first edition, the famous geometer Gian-Carlo Rota wrote [Rota, 1999, 19]:

> Today, synthetic geometry is still the downside. For today's students of algebraic geometry, points, lines and surfaces are a manner of speaking, shorthand terms for algebraic concepts. But the call to reality is making itself felt. Computer scientists have shown us how little we know about solid angles, how much we need to know about polyhedra. The visual geometry of Euclid, Desargues, Ludwig Schläfli and Eugenio Cremona is about to make a triumphal comeback. Geometers of today will be well advised to recover their bearings by reading Hilbert's *Grundlagen der Geometrie*.

BIBLIOGRAPHY

Blumenthal, O. 1935. 'Lebensgeschichte', in [Hilbert, *Papers*], vol. 3, 388–429.
Contro, W. 1976. 'Von Pasch zu Hilbert', *Archive for history of exact sciences, 15*, 283–295.
Dehn, M. 1922. 'Hilberts geometrisches Werk', *Die Naturwissenschaften, 10*, 77–82.

Fano, G. 1892. 'Sui postulati fondamentali della geometria proiettiva', *Giornale di matematiche, 30,* 106–132.

Freudenthal, H. 1957. 'Zur Geschichte der Grundlagen der Geometrie', *Nieuw Archief voor Wiskunde, (4) 5*, 105–142.

Hilbert, D. *Papers*. *Gesammelte Abhandlungen*, 3 vols. Berlin: Springer, 1932, 1933, 1935. [Repr. New York: Chelsea, 1965.]

Hilbert, D. and Cohn-Vossen, S. 1932. *Anschauliche Geometrie*, Berlin: Springer. [Repr. Darmstadt: Wissenschaftliche Buchgesellschaft, 1973. English trans.: *Geometry and the imagination*, New York: Chelsea, 1952.]

Klein, F. 1914. *Elementarmathematik vom höheren Standpunkte aus*. *Teil II: Geometrie*, 2nd ed., Leipzig: Teubner.

Pasch, M. 1882. *Vorlesungen ueber neuere Geometrie*, 1st ed., Leipzig: Teubner. [2nd ed. 1926, repr. 1976.]

Peano, G. 1891. *Die Grundzüge des geometrischen Kalküls*, Leipzig: Teubner. [Original ed.: *Calcolo geometrico secondo l'Ausdehnungslehre di H. Grassmann*, Turin: Bocca, 1888.]

Peano, G. 1889. *I principii di geometria logicamente esposti*, Turin: Bocca. [Repr. in *Opere scelte*, vol. 2, 56–91.]

Reid, C. 1970. *Hilbert*, New York and Berlin: Springer.

Rota, G.-C. 1999. 'Geometrie heute—Eine Umfrage', *Mitteilungen der Deutschen Mathematiker-Vereinigung, H.1,* 17–20.

Toepell, M. 1985. 'Zur Schlüsselrolle Friedrich Schurs bei der Entstehung von David Hilberts "Grundlagen der Geometrie' ", in *Mathemata. Festschrift für Helmuth Gericke*, Stuttgart: Franz Steiner, 637–649.

Toepell, M. 1986a. *Über die Entstehung von David Hilberts, Grundlagen der Geometrie*', Goettingen: Vandenhoeck & Ruprecht.

Toepell, M. 1986b. 'On the origins of David Hilbert's "Grundlagen der Geometrie" ', *Archive for history of exact sciences, 35*, 329–344.

Toepell, M. 1999a. 'Zur Entstehung und Weiterentwicklung von David Hilberts "Grundlagen der Geometrie"', in Hilbert, *Grundlagen der Geometrie,* 14th ed., Leipzig: Teubner, 283–324.

Toepell, M. 1999b. 'Die projektive Geometrie als Forschungsgrundlage David Hilberts', in *ibidem*, 347–361.

Weyl, H. 1944. 'David Hilbert and his mathematical work', *Bulletin of the American Mathematical Society, 50,* 612–654. [Shortened version in [Reid, 1970, 245–283].]

Wiener, H. 1891. 'Über Grundlagen und Aufbau der Geometrie', *Jahresbericht der Deutschen Mathematiker-Vereinigung, 1*, 45–48.

CHAPTER 56

KARL PEARSON, PAPER ON THE CHI SQUARE GOODNESS OF FIT TEST (1900)

M.E. Magnello

This paper contains the first mathematical account for a goodness of fit test that could be used for any shape curve including, for example, Poisson, binomial and Mendelian distributions, rather than simply the normal distribution. Together with other papers of the time by Pearson and colleagues, it raised substantially the practice of mathematical statistics.

First publication. 'On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling', *Philosophical magazine, (5) 50* (1900), 157–175; correction in *(6) 1* (1901), 670–671.

Reprint. In Pearson, *Early statistical papers*, Cambridge: Cambridge University Press, 1948 (repr. 1956), 339–357.

Related articles: Laplace (§24), Fisher (§67), Shewhart (§72).

1 EDUCATION AND EMPLOYMENT

The education of Karl Pearson (1857–1936) began when he had French lessons at the age of four at his home on the Camden Road, London. A couple of years later he and his brother Arthur went to a small school in Harrow established by a William Penn; eventually, the boys attended University College London School on Gower Street. When Karl was 15 years old, his father was looking for a good Cambridge Wrangler to prepare him for the Mathematics Tripos. Less than a year later, Karl went up to Hitchin, where he stayed for five months receiving tuition from the Reverend Louis Hensley. Very unhappy there, he left on 1 July 1874 to go to Merton Hall, Cambridge to be coached in mathematics under John Edward Rendall Harris, John P. Taylor and the great mathematics Cambridge tutor John Edward Routh. He stayed for a year from mid July 1874 until 15 April 1875, when he received an Open Fellowship from Kings College, Cambridge.

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

Though Pearson was not a very healthy child, he came to life in this environment and his health improved: he found the highly competitive and demanding system leading up to the Mathematical Tripos was the tonic he needed. His tutors included Arthur Cayley, William Herrick Macaulay, George Stokes, Percival Frost, Isaac Todhunter and James Clerk Maxwell. Pearson graduated with honours in January 1879, being the Third Wrangler in the Mathematics Tripos; subsequently, he received a Fellowship from King's College, which gave him financial independence for seven years, and was made an Honorary Fellow in 1903. Immediately after he graduated, he read philosophy and medieval languages before going to Germany where he studied physics and metaphysics under Kuno Fischer, Hermann von Helmholtz, Gustav Kirchhoff and Heinrich Quincke in Heidelberg, and then the law in Berlin.

Between 1879 and 1884 Pearson applied for mathematical posts in Dundee, Leeds, Manchester and Sheffield. He read law at Lincoln's Inn and was called to the Bar in 1881, but practiced for only a very short time. He took on a temporary job teaching mathematics at King's College, London in 1883 when the professor fell ill. Then June 1884 he accepted the Goldsmid Chair of 'Mechanism and Applied Mathematics' at University College London (hereafter, 'UCL'), succeeding the German-born mathematician Olaus Henrici (1840–1918). Henrici's predecessor was the philosopher and mathematician William Kingdon Clifford (1845–1879), whose book, the *Common sense of the exact sciences,* Pearson was asked to finish after Clifford's early death; it appeared in 1885. During Pearson's first six years at UCL he taught modern geometry, mathematical physics, statics, dynamics, hydrodynamics, magnetism, sound, electricity and elasticity theory (his speciality) to engineering students. As well as publishing several research papers, he completed Todhunter's *A history of the theory of elasticity and the strength of materials* (1886–1893).

Being ambitious, Pearson also took up the Gresham Chair of Geometry at Gresham College in the City of London in 1890, and held it for three years, concurrently with his post at UCL. He delivered a total of 38 lectures between the spring of 1891 and the summer of 1894. These Gresham lectures signified a turning point in his career when he began to teach the geometry of statistics and especially when he helped the Darwinian zoologist W.F.R. Weldon (1860–1906) with his data on marine organisms [Magnello, 1996].

2 CHRONOLOGY AND CURVE FITTING

The development of Pearson's work on curve-fitting and finding a goodness of fit test for asymmetrical distributions may be seen as his reaction to the status accorded to the normal distribution at the end of the 19th century. The idea that empirical data should conform to the normal distribution was borne out of the use by Adolphe Quetelet (1796–1874) of the arithmetic mean for his *l'homme moyen*: Quetelet's deterministic outlook led him to believe that deviations from the ideal type were flawed and thus a product of error. This interpretation of error arose from the theistic argument that order and perfection of nature in the universe were due to the existence of a divine Creator (i.e., 'the argument from design'). Thus, any error that arose in nature could only be viewed as flawed, as it would have interfered with the Deity's plan and purpose of the universe. In fact, Quetelet had made one of the earliest attempts to fit a set of observational data to a normal curve in 1840 that

Francis Galton (1822–1911) began to use in 1863. Wilhelm Lexis (1837–1914) devised the Lexican Ratio L as a goodness of fit test to determine if an empirical distribution conformed to the normal distribution [Lexis, 1877], whilst Francis Ysidro Edgeworth provided a goodness of fit test that was based on a normal approximation to the binomial distribution [Edgeworth, 1885]. Though many other 19th-century scientists attempted to find a goodness of fit test, such as the American statistician Erasmus Lyman de Forest and the Italian Luigi Perozzo, they did not give any underlying theoretical basis for their formulas.

Pearson's interest in curve fitting was fuelled by Weldon's work on the Plymouth shore crabs when Weldon found that one of his distributions of data was asymmetrical, while the rest of his data were normally distributed. Since Weldon's data produced a bimodal curve instead of the normal curve, Pearson wanted to find another way to interpret the data without trying to normalise it as Quetelet and Galton had done. Pearson and Weldon thought it was important to make sense of the shape of the distribution without distorting its original shape, as it might have revealed something about the creation of new species. Pearson adapted the mathematics of mechanics, using the *method of moments*, to construct a new statistical system to interpret Weldon's data since no such system existed at the time. This system allowed Pearson to analyze data of all kinds of shapes, and enabled him to move beyond the limitations of the normal curve.

Beginning in 1892, Pearson deployed higher moments, that is, integrals of the form $\int f(x)x^n\,dx$ for $n > 2$, where f is some function and x a distance measured from a given point. These moments had been used in statistics around the 1850s by Jules Bienaymé (compare section 3 below) and Pafnuty Chebyshev, and 20 years later by A.A. Markov; but Pearson learnt of them from graphical statics, when determining moments on a loaded continuous beam with the help of a theorem proved in 1857 by the French engineer Emile Clapeyron [Pearson, 1890].

In mechanics, the moment of a force about a point, such as a fulcrum, is the product of the magnitude of that force by its perpendicular distance from that point. In statistics, moments are averages, and force was replaced by a frequency curve function (such as the percentage of the distribution within a given class interval).

In his Gresham lecture on statistics [Pearson, 1893] showed that the first moment of any set of lines at unit distance from each other is the sum of their lengths multiplied by their respective distances from a parallel straight line about which he found the moment—that is, it is the mean. The second moment is the sum of their lengths multiplied by the squares of their distances, the square of the standard deviation: Pearson called it the 'squared standard deviation', which R.A. Fisher (1890–1962) termed the 'variance' in 1918. The third moment is the sum of their lengths multiplied by the cubes of their distances, and is used to find a measure of skewness of a distribution. The fourth moment is found by multiplying the lengths by the fourth power. The fourth moment measures how flat or peaked is the curve of the distribution; for it he coined the word 'kurtosis' (from the Greek word for bulginess), It had three further components: (a) if data clustered or peaked around the mean (he called the peakedness 'leptokurtic'); or (b) if it spread out across the distribution, the curve 'platykurtic', for it resembled the shape of a platypus; or (c) if it produced a normal curve, this was termed 'mesokurtic'.

After Pearson examined Weldon's asymmetric curves derived from his crab data in Naples, he realised that an objective method of measuring the goodness of fit was a desider-

atum for distributions that did not conform to the normal curve. Pearson's earliest consideration of determining a measure of the goodness of fit test came out of his lecture on 21 November 1893 when he introduced the sixth moment as a measure of a goodness of fit. At the start of 1894 he produced the forerunner to his chi-square goodness of fit test. He demonstrated how to find $\sum \mathbf{s}/\mathbf{y}$, where **s** equalled the difference between the observation polygon and the theoretically expected curve and **y** its corresponding ordinate. Since Pearson thought this measure 'was awkward to get', he decided that it would be preferable to measure the ratio of the whole area between the theoretically expect curve and the polygon (of observational data); he counted all the value as positive which equalled W and then measure the total area A under the curve. Hence

$$W/A = \Big(\sum \text{errors of fit}\Big)/\Big(\sum \text{ordinates}\Big) = \sum s/\sum y, \tag{1}$$

which he thought was a reasonably measure of the goodness of fit. On Christmas Day in 1896, he wrote to Galton that he wanted to develop a goodness of fit test for asymmetrical distributions for biologists and economists. He reached a resolution in 1900, when he found the exact chi-square distribution from the family of Gamma distributions and devised his *chi-square* (χ^2, P) *goodness of fit test*.

For Pearson, the object of the chi-square goodness of fit test was to find 'a criterion of the probability on any theory of an observed system of errors, and apply it to the determination of goodness of fit in the case of frequency curves'. This was to be used to determine how well an observed or empirical distribution could be fitted to a theoretical frequency distribution. Prior to this test, the usual procedure involved comparing errors of observation, to a table of distributions based on the normal curve, or graphically by means of a plotted frequency diagram. From the basis of these comparisons, error theorists maintained that 'an experimental foundation has been established for the normal law of errors [...] having deduced the normal curve of errors, they gave as a rule some meagre data of how it fits the observation' (p. 171).

The statistical framework of Pearson's chi-square is a tripartite system, for it not only incorporates a probability distribution and a goodness of fit test, but it also includes a statistical technique that Pearson introduced in [1904] when he extended his goodness of fit test to manifold classifications for the analysis of contingency tables. He termed this technique the 'chi-square contingency coefficient' to test differences between observed cell frequencies and theoretically expected cell frequencies. Fisher [1922] renamed it the *chi-square statistic* (§67).

The chi-square goodness of fit test and the chi-square test of association for contingency tables are built on two different hypotheses. The hypothesis of the goodness of fit test of the 1900 paper tests a hypothesis that the relative frequencies of mutually exclusive observed events follow a specified frequency distribution, and seeks to determine if the observed distribution (constructed from observational data) conforms to the theoretical distribution, which could be the normal, the Poisson, binomial or Mendelian distributions. The chi-square test of association in [Pearson, 1904] seeks instead to determine if an association exists between two discrete variables (such as Mendelian alleles) in a contingency table.

3 THE MATHEMATICAL DERIVATION

The derivation of the chi-square distribution has its roots in the *method of least squares*, which had been derived from the theory of errors by astronomers in the middle of the 19th century. The astronomers arrived at the method of least squares when they realised they needed to choose from alternative estimators (that is, other measures of dispersion) and they came up with the first deviation of the distribution of a sum of squares $\sum(\bar{X} - X)^2$ or $\sum x^2$: this provided the mathematical basis for the chi-square distribution. The chi-square distribution is produced by generating a random collection of a series of these square deviational values (that is, $\sum x^2$).

Pearson's chi-square distribution can be regarded as the culmination of the least squares theory applied to discrete distributions. Gamma distributions were used by Pierre Simon Laplace for his work in error theory in the early 19th century (§24). Though Laplace had not obtained the distribution of the sum of squares nor the usual gamma distributions for the continuous chi-square distribution, he found instead a distribution of the precision underlying a Bayesian hypothesis gave the observations. Moreover, Pearson had, after the event, 'provided all the necessary mathematical techniques for Bienaymé to obtain the distribution of chi-square as an asymptotic result without the assumption of normality' [Lancaster, 1966].

Bienaymé's extension of the work of Laplace in his 1838 Bayesian study of the lunar form of multivariate variables has been taken to show that 'he had very nearly anticipated Pearson's work on the normal approximation to the multinomial' [Lancaster, 1966]. Fourteen years after Bienaymé extended Laplace's work, he used the gamma distribution of the sum of square in the least squares theory. When Pearson derived the sampling distribution of the χ^2 in large samples, he found that it was a specialised form of his Type III distribution that he derived in 1896. Both of these distributions are positively skewed, unless n increases and then the χ^2 distribution will approach the normal distribution as its limit.

Thus $\chi^2 =$ constant is the equation of a generalised ellipsoid, and is the surface of a frequency of the system of errors or deviations $x_1, x_2, \ldots, x_n$. The value given to χ, which covered the whole of the generalised space, ranged from 0 to ∞. As soon as the observed deviational value, the standard deviation and the correlation of errors were known, the value of χ could be found. Pearson then determined the probability (P) of the ellipsoid and found a value of P for a series of values of χ^2 for different cases. The value of P could be determined once χ^2 had been calculated and when the sample size or 'number of independent observations' (or what is termed today *degrees of freedom*) $n' = n - 1$ was found.

Pearson first considered cases in which the theoretical probability is known *a priori*. He took

$$\chi^2 = S\frac{(m' - m_s - \mu)^2}{m_s}, \tag{2}$$

where $m' =$ observed (or empirical) frequencies in a distribution, $m_s =$ theoretical (or expected) a distribution known *a priori*, μ the population mean, and S marks summation.

A more contemporary formula for the chi-square goodness of fit test is

$$\chi^2 = \sum (O - E)^2 / E \quad \text{for observed values } O \text{ and expected values } E. \tag{3}$$

However, when the theoretical distribution had to be judged from the sample itself, it must be determined if the sample represents a random system of deviations from the theoretical frequency distribution of the general population, but this distribution has to be inferred from the sample itself. Pearson thus modified the test for distributions not known *a priori*.

4 SOME LATER DEVELOPMENTS

4.1 The founding of Biometrika. With the P, χ^2 measure, it was then possible to determine whether an empirical frequency curve could describe effectively the sample drawn from the given population (that is, the theoretical curve). If it were a bad fit, then this curve could be used to describe other samples from the same population and when the value for the χ^2 test increases, the fit becomes worse. Whilst Pearson used the chi-square goodness of fit test initially for games of chance in a binomial distribution and for data from Hugo de Vries's buttercups for asymmetrical distributions, Weldon was to make the earliest use of the chi-square goodness of fit test for a Mendelian distribution. The ensuing and sometime vitriolic debates between Pearson and Weldon and the early Mendelians, especially William Bateson, about using statistical methods for problems of biology, led them to establish the journal *Biometrika* in November 1900, seven months after the landmark paper was published. Weldon then published there a paper [Weldon, 1902] on using the chi-square goodness of fit test on a Mendelian distribution using Mendel's data in the second volume of *Biometrika*. Pearson [1927] later used the chi-square goodness of fit test for Poisson distributions [Magnello, 1998].

4.2 Minimum chi-squared. After Pearson introduced the chi-square goodness of fit test in 1900, several authors tried basing estimation on χ^2 (see especially [Engledow and Yule, 1914]). Two years later the biometrician Kirstine Smith first used the phrase 'minimum χ^2', though only in tables where brevity was necessary [Smith, 1916]. Fisher also used the minimum χ^2 for he often compared the method with his own maximum likelihood.

4.3 Yates's chi-square correction for continuity. This adjusts the formula for Pearson's chi-square test for small sample sizes by subtracting 0.5 from each observed value in a 2×2 contingency table [Yates, 1934]. This formula is mainly used when at least one cell of the table has an expected frequency less than 5.

5 CONCLUDING REMARKS

Pearson's work on curve fitting meant that he needed a criterion to determine how good the fit was, which led him to devise different goodness of fit tests. This idea underpinned the infrastructure to his statistical theory and encompassed his entire working life as a statistician; it began in 1892 when he introduced the sixth moment as a measure of a goodness of fit for Weldon's crab data, continued throughout the 1890s, culminated with this

1900 landmark paper when he devised the chi-square goodness of fit test, and ended with the last paper, written when he was 79 years old [Pearson, 1936]. This paper, published posthumously two months after his death, was written in response to [Fisher, 1922] on maximum likelihood as a means of curve fitting that he had introduced in 1921. Though Fisher regarded Pearson's chi-square goodness of fit test as his most important contribution to statistics, he challenged Pearson's method of moments system for curve fitting when he introduced maximum likelihood. The likelihood of a parameter is proportional to the probability of the data; as a function it usually has a single maximum value, which Fisher called the 'maximum of likelihood'.

Throughout the 20th century Pearsonian statistics exerted a major influence on the development of various disciplines and continues to play a pivotal role in industry, education and in the biological, behavioural, medical, social and human sciences. His statistical methods transformed biological and medical statistics in the 20th century, especially owing to his students Major Greenwood and Austin Bradford Hill; for example, the latter went on to create the first randomised clinical trials in modern therapeutic medicine. Another student, Charles Spearman, was also influenced by Galton's ideas of measuring individual differences in human abilities; his early ideas on intelligence testing used Pearson's product-moment correlation and method of principal components to create a new statistical method, known as 'factor analysis', which reduces a set of complex data into a more manageable form and makes it possible to detect structures in the relationship between variables. With this new tool, Spearman went on to create the first psychometric theory of intelligence with his two-factor theory, which measured general and specific abilities. Other psychometricians devised a battery of aptitude, psychological and other psychometric tests.

The first statistical quality control test for industry was devised by Pearson's student William Sealy Gosset, who used the pseudonym 'Student'; his work inspired Fisher to create a statistical system for the analysis of small samples (that is, analysis of variance), thereby introducing experimental design and randomisation into statistical theory. Fisher's statistical innovations inaugurated the second phase in the development of modern mathematical statistics through his development of inferential statistics; the distinctive feature of the newer form of statistics involved the formal testing of hypotheses and parameter estimation by using properties of consistency, unbiasedness, efficiency and sufficiency. The foundations of Fisher's method had not only been built upon Pearson's statistical work, but Fisher's [1922] paper represented a translation of Pearson's statistical language and became the vernacular of contemporary mathematical-statistical theory (§67), even though many of Pearson's statistical methods and his language remain a part of contemporary statistical theory.

BIBLIOGRAPHY

Aldrich, J. 2003. 'The language of the English biometric school', *International statistics review, 71*, 109–129.

Edgeworth, F.Y. 1885. 'The empirical proof of the law of errors', *Philosophical magazine, (5) 24*, 330–347.

Edwards, A.W.F. 1997. 'Three early papers on efficient parametric estimation', *Statistical science, 12*, 35–38.

Elderton, W.P. 1901. 'Tables for testing the goodness of fit', *Biometrika, 1*, 159.

Engledow, F.L. and Yule, G.U. 1914. 'The determination of the best value of the coupling-ratio from a given set of data', *Proceedings of the Cambridge Philosophical Society, 17*, 436–440.

Fisher, R.L. 1922. 'On the mathematical foundations of theoretical statistics', *Philosophical Transactions of the Royal Society of London, (A) 222*, 309–368.

Lancaster, H. 1966. 'Forerunners of the Pearson χ^2', *The Australian journal of statistics, 8*, 117–126.

Lexis, W. 1877. *Zur Theorie der Massenerscheinungen in der menschlichen Gesellschaft*, Freiburg in Breisgau: Wagner.

Magnello, M.E. 1996. 'Karl Pearson's Gresham lectures: W.F.R. Weldon, speciation and the origins of Pearsonian statistics', *British journal of the history of science, 29*, 43–63.

Magnello, M.E. 1998. 'Karl Pearson's mathematisation of inheritance: From Galton's ancestral heredity to Mendelian genetics', *Annals of science, 55*, 35–94.

Magnello, M.E. 2005. *Victorian values. The origin of modern statistics*, Cambridge: Icon Press.

Pearson, E. 1965. 'Early history of biometry and statistics 1890–94', *Biometrika, 52*, 3–18.

Pearson, K. 1890. 'Note on Clapeyron's theorem of the three moments', *The messenger of mathematics, 19*, 129–135.

Pearson, K. 1893. Gresham lecture on 'Skew curves', 23 November 1893. Pearson Papers, Manuscript Room, University College London.

Pearson, K. 1904. 'Mathematical contributions to the theory of evolution. XIII. On the theory of contingency and its relation to association and normal correlation', *Drapers' Company Research Memoirs. Biometric series, I* (London: Dulau), 1–37.

Pearson, K. 1911. 'On the probability that two independent distributions of frequency are really samples of the same population', *Biometrika, 8*, 250–254.

Pearson, K. 1916. 'On the general theory of multiple contingency with special reference to partial contingency', *Nature, 11*, 159.

Pearson, K. 1927. 'Note on the relation of the (χ^2, P) goodness of fit test to distributions of standard deviation in samples from a normal population' *Biometrika, 19*, 215.

Pearson, K. 1936. 'Method of moments and the method of maximum likelihood', *Biometrika, 28*, 34–59.

Plackett, R. 1983. 'Karl Pearson and the chi-squared test', *International statistical review, 51*, 59–72.

Smith, K. 1916. 'On the "best" values of the constants in frequency distributions', *Biometrika, 11*, 262–276.

Stigler, Stephen M. 1986. *The history of statistics*: *the measurement of uncertainty before 1900*, Cambridge, MA: The Belknap Press.

Weldon, W.F.R. 1902. 'Mendel's law of inheritance in peas', *Biometrika, 1*, 109–124.

Yates, F. 1934. 'Contingency tables involving small numbers and the χ^2 test', *Supplement to journal of the Royal Statistical Society*, 1, 217–235.

CHAPTER 57

DAVID HILBERT, PAPER ON 'MATHEMATICAL PROBLEMS' (1901)

Michiel Hazewinkel

In this remarkable paper, based upon a lecture delivered to the International Congress of Mathematicians in Paris in 1900, Hilbert outlined a range of problems for mathematicians to address in the century about to start. Indeed, they were to have a marked influence on the development of several branches of mathematics.

First publication. 'Mathematische Probleme', *Nachrichten der Königlichen Gesellschaft der Wissenschaften zu Göttingen, mathematisch-physikalische Klasse,* (1901), 253–297. Also in *Archiv der Mathematik und Physik, (3) 1* (1901), 44–63, 213–237 [slightly revised].

Reprints. In *Gesammelte Abhandlungen,* vol. 3, Berlin: Springer, 1935, 290–309. Also in P.S. Alexandrov (ed.), *Die Hilbertschen Probleme,* Leipzig: Geest und Portig, 1971, 22–80. Also in R. Bellman (ed.), *A collection of modern mathematical classics. Analysis,* New York: Dover, 1961, 247–292.

Partial French translations. 1) In *L'enseignement mathématique, (1) 2* (1900), 349–355. 2) In *Revue générale des sciences pures et appliquées, 12* (1901), 168–174.

Full French translation. As 'Sur les problèmes futurs des mathématiques', in E. Duporcq (ed.), *Compte rendu du Deuxième Congrès International de Mathématiciens*, Paris: Gauthiers–Villars, 1902 (repr. Liechtenstein: Kraus, 1967), 58–114. Also published separately.

English translation by M. Newsom in *Bulletin of the American Mathematical Society, 8* (1902), 437–479. [Repr. in *new series, 8* (2000), 407–436. Also in F.E. Browder (ed.), *Mathematical developments arising from Hilbert problems,* 2 vols., Providence: American Mathematical Society, 1976, 1–34. Also available at http://aleph0.clarku.edu/~djoyce/hilbert.]

Russian translation. In P.S. Alexandrov (ed.), *Problemi Gilberta,* Moscow: Nauka, 1969.

Related articles: many in this volume.

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

1 INTRODUCTION

In August 1900, at the occasion of the second International Congress of Mathematicians in Paris, David Hilbert (1862–1943), then all of 38 years young, gave his lecture on 'mathematical problems'. That lecture and even more the written version of it has been of great influence on the development of mathematics in the 20th century, or so it would seem. It stems partly because of the stature of the lecturer, which was still to grow considerably in the decades to come; partly because the problems were well chosen; partly because they breathed a coherent view of what mathematics is all about; and perhaps most of all because of the incurable optimism in it all, a flat denial of Emil Du Bois-Reymond's claim 'Ignoramus et ignoramibus'.

The full published version (see above) contains 23 problems. Of these Hilbert discussed only 10 in the lecture itself (numbers 1, 2, 6, 7, 8, 13, 16, 19, 21, 22). The 23 problems, together with short, mainly bibliographical comments, are surveyed below using the short title descriptions from the full versions.

Three general references are [Alexandrov, 1979] for all 23 problems; [Browder, 1976] for all problems except 2, 3 and 16; and [Kantor, 1996] for all problems except 4, 9 and 14, and with special emphasis on developments from 1975 to 1992. Two semipopular accounts of the problems, their solutions or solution attempts, and the people who worked on them are [Gray, 2000] and [Yandell, 2002]. The account below is mostly based on [Hazewinkel, 2000], and the references quoted there.

2 THE PROBLEMS

PROBLEM 1. *Cantor's problem on the cardinal number of the continuum.*

More colloquially also known as the *continuum hypothesis*, it can be stated as 'Every uncountable subset of the real numbers, **R**, has the same cardinality as **R**', or as the statement '$2^{\aleph_0} = \aleph_1$'.

It was solved by Gödel [1939] and Cohen [1963], in the (unexpected) sense that the continuum hypothesis is independent of ZFC, the *Zermelo–Frankel axioms* of set theory complete with the axiom of choice. This means that one can add the continuum hypothesis to ZFC without introducing inconsistencies (that were not already present) (Gödel); one can also add the negation of the continuum hypothesis (Cohen) without introducing inconsistencies. Gödel and Cohen also showed that the axiom of choice is independent of ZF.

Perhaps even more important than the solution of the problem itself are the techniques of Cohen forcing and Boolean-valued models that resulted. These have 'uncountably' many applications by now.

PROBLEM 2. *The compatibility of the arithmetical axioms.*

This was solved (in a negative sense) by Gödel [1931] with the so-called 'Gödel incompleteness theorem'. It roughly says that in every system that is strong enough to do a reasonable amount of arithmetic there are statements that are not provable within that system

and whose negation is also not provable. For a popular account, see [Nagel and Newman, 1959]. Positive results (using techniques that Hilbert would not have allowed) are due to G. Gentzen in 1936 and P. S. Novikov in 1941 (A.S. Essenin-Vol'pin in [Alexandrov, 1979], G. Kreisel in [Browder, 1976]).

PROBLEM 3. *The equality of the volumes of two tetrahedra of equal bases and equal altitudes.*

More precisely, the problem was to show that two such polyhedra can be transformed into each other by cutting and pasting (as is the case for triangles, the analogous problem in dimension 2). This is the origin of the name 'scissors congruence problems'.

It was solved in the negative sense by Hilbert's student Max Dehn in 1900, actually before Hilbert's lecture was delivered [Dehn, 1901], and at least partially already in [Bricard, 1896]. As it turned out, there is besides the volume one more quantity that remains invariant under cutting and pasting, the *Dehn invariant*. In higher dimensions the same problem can be studied and there are the *Hadwiger invariants*. In dimension 3 the Dehn invariant is the only extra invariant besides volume, i.e. tetrahedra with the same Dehn invariant and the same volume are scissors congruent; this is due to J.P. Sydler in 1965 [Sah, 1979].

PROBLEM 4. *Problem of the straight line as the shortest distance between two points.*

This problem asks for the construction of all metrics in which the usual lines of projective space (or pieces of them) are geodesics. The first work on this was by Hilbert's student G. Hamel in [Hamel, 1903]. In particular, he pointed out that the problem needed to be made more precise, and that one should ask for all *Desarguesian* spaces in which straight lines are the shortest distances between points. Nowadays, the problem is considered (basically) solved in the form of the following (generalized) Pogorelov theorem: Any n-dimensional Desarguesian space of class C^{n+2}, $n \geqslant 2$, can be obtained by the BB construction, that is, a technique based upon integral geometry for obtaining Desarguesian spaces due to Blaschke [1936] and Busemann [1961]. The differentiability class restriction is necessary, for otherwise there are Desarguesian spaces that do not come from the BB construction; see [Szabo, 1986], which is also recommended as a very good survey of the fourth problem, and also [Pogorelov, 1973, 1979].

PROBLEM 5. *Sophus Lie's concept of a continuous group of transformations without the assumption of the differentiability of the functions defining the group.*

This was solved in [Gleason, 1952] and [Montgomery and Zippin, 1952], in the form of the theorem 'Every locally Euclidean topological group is a Lie group and even a real analytic group'. For a much simplified but non-standard treatment see [Hirschfield, 1990]. The cases of compact topological groups and commutative topological groups were handled earlier by J. von Neumann in 1933 and L.S. Pontryagin in 1934.

This is perhaps the only one of Hilbert's problems that did not give rise to a host of subsequent investigations and problems and concepts. This happens but rarely. As M. Davis writes in his discussion of the first problem in [Browder, 1976], after Gödel's work there

was some 20 years of stagnation in set theory; but this period served to set people thinking about computability, recursiveness and the like, a most important development that prepared ground for modern computer science and vast new parts of logic.

PROBLEM 6. *Mathematical treatment of the axioms of physics.*

This is very far from solved in any way, though there are many (bits and pieces of) axiom systems that have been investigated in depth. A.S. Wightman has given an extensive discussion of Hilbert's own ideas, von Neumann's work and much more in [Browder, 1976]. There are, for instance, the *Wightman axioms* (also called *Gärding–Wightman axioms*) and the *Osterwalder–Schrader axioms* of quantum field theory; and von Neumann's axiomatization of quantum mechanics [von Neumann, 1932] (§69), following work of P. Nordheim and Hilbert himself. More recently there is the definition of topological field theories and conformal field theories, sources of very fruitful interactions between mathematics and physics [Lawrence, 1996; Sawin, 1996; Segal, 1988, 1991; Turaev, 1994, ch. 2; Witten, 1988]. Note that these are not really axiomatizations from the ground up (like Euclidean geometry) but are more aptly termed 'relative axiomatizations' in that they take an existing body of knowledge (like, say, differential topology) as given.

Quite early in the 20th century, in direct response to Hilbert's questions, there were Hamel's axiomatization of mechanics in 1903, Constantin Carathéodory's axiomatizations of thermodynamics in 1909 and of special relativity in 1924, and another independent axiomatization of special relativity by A.A. Robb in 1914. Finally, there was R. von Mises's axiomatization of probability (a field specifically mentioned by Hilbert in his elucidation of problem 6) followed by the definitive axiomatization by Kolmogorov [1933] (§75).

Preliminary to the axiomatization of quantum mechanics there was the development of Hilbert space, operators, infinite matrices, eigenvalues and integral equations. Hilbert remarked that he developed this theory on purely mathematical grounds and even called it 'spectral analysis' without any idea that it would later be much related to the real spectra of physics [Reid, 1970, p. 183].

PROBLEM 7. *Irrationality and transcendence of certain numbers.*

The numbers in question are of the form α^β with α algebraic and β algebraic and irrational; for instance $2^{\sqrt{2}}$ and $e^\pi = i^{-2i}$. The problem was solved in 1934 by A.O. Gel'fond and Th. Schneider (the *Gel'fond–Schneider theorem*). For the general method, the *Gel'fond–Baker method*, see R. Tijdeman in [Browder, 1976]. A large part of [Fel'dman and Nestorenko, 1998] is devoted to this problem and related questions.

It is interesting to note that in a lecture given in 1919 Hilbert remarked that he was optimistic to see the Riemann hypothesis solved in his lifetime, that perhaps the youngest member in the audience would see the solution of the Fermat problem, but that no one in the audience would see the transcendence of $2^{\sqrt{2}}$ [Gray, 2000].

PROBLEM 8. *Problems of prime numbers.*

This one is usually known as the *Riemann hypothesis* and is the most famous and important of the unsolved conjectures in mathematics. The Riemann zeta-function of the complex variable s is given for $\mathrm{Re}(s) > 1$ by $\zeta(s) = \sum_{i=1}^{\infty} n^{-s}$ and it has an analytic continuation to the whole s-plane to a meromorphic function with one simple pole at $s = -1$ with residue 1; and zeros for $s = -2, -4, \ldots$, referred to as 'trivial zeros'. The Riemann hypothesis now says that all other zeros are of the form $1/2 + i\tau$ (using the Gauss plane notation for a complex number). It is known that the first 1.5 billion zeroes (arranged by increasing positive imaginary parts) are simple and lie on the critical line $\mathrm{Re}(s) = 1/2$ [van de Lune et al., 1986]; also that more than 40% of the zeros satisfy the Riemann hypothesis [Selberg, 1942; Levinson, 1974; Conrey, 1989].

The zeta function in algebraic geometry, $\zeta_X(s)$, is a meromorphic function of a complex variable s that describes the arithmetic of algebraic varieties X over finite fields or of schemes of finite type over the integers. If X is $Spec(\mathbf{Z})$, then one recovers the Riemann zeta function; if X is of finite type over $Spec(\mathbf{Z})$, then there result the Dedekind zeta functions for the corresponding number fields.

André Weil formulated a number of far-ranging conjectures concerning zeta functions of varieties over finite fields, and proved them for curves. After the necessary cohomological tools for this were developed by A. Grothendieck (mostly), M. Artin and J.-L. Verdier, these conjectures were proved by [Deligne, 1974, 1980]; for details see [Parshin, 1993].

PROBLEM 9. *Proof of the most general law of reciprocity in any number field.*

Consider the question of whether an integer a is a quadratic residue modulo a prime number p or not, where a is not divisible by p. I.e. the question is whether a can be written in the form $(b^2 + kp)$ for some integers b and k or not. In the first case write $(\frac{a}{p}) = 1$, in the second $(\frac{a}{p}) = -1$. This is the definition of the Legendre symbol. The Gauss reciprocity theorem now says that for two different odd prime numbers $(\frac{p}{q})(\frac{p}{q}) = (-1)^{\frac{p-1}{2}\frac{q-1}{2}}$.

In 1927 Artin [1928] gave reciprocity laws for general number fields. A great generalization of the Gauss reciprocity law had already been established by Hilbert himself in 1895 and 1896; see [Stepanov, 1992] for more details and also for information as to how the question of reciprocity laws leads to (Abelian) class field theory, the subject of problem 12 below. The analogous question of reciprocity laws for function fields was settled in [Shafarevich, 1950] as the *Shafarevich reciprocity law*.

PROBLEM 10. *Determination of the solvability of a Diophantine equation.*

A Diophantine equation in a finite number of variables is an equation $P(x_1, \ldots, x_n)$ where P is a polynomial over the integers. It is solvable if there are integral solutions. For instance, the Fermat equation $x^n + y^n = z^n$ for a given natural number n which has infinitely many solutions for $n = 1$ and 2 and no solutions for larger n. (Hilbert referred to this problem in the preamble of his paper.) The problem asks for a finite sequence of tests

(that can be applied to any such equation) to determine whether a Diophantine equation has solutions or not.

The solution is negative: Yu. Matiyasevich showed in 1970 that there is no such algorithm. This is a fairly immediate consequence of the main theorem in the field: Every listable set of natural numbers is Diophantine. For a description of the various concepts (though the meaning is intuitively rather clear), see Davis and others in [Browder, 1976].

One consequence of the main theorem is that there exists an integral polynomial such that the positive values of this polynomial on the natural numbers are precisely the prime numbers [Putnam, 1960]. This result made many mathematicians doubt that the main theorem, at that time still a conjecture, could possibly be true. For a discussion of various refinements and extensions of the problem, see [Pheidas, 1994].

PROBLEM 11. *Quadratic forms with any algebraic numerical coefficients.*

This problem asks for the classification of quadratic forms over algebraic number fields. More precisely, a quadratic form over a (number) field K is an expression of the form $\sum_{i\leqslant j} q_{I,j}x_i x_j$ in the variables $x_1,\ldots,x_n$ with coefficients in K. Two such forms q and q' are equivalent if there is an invertible linear substitution $x_i' = \sum_{i\leqslant j} t_{i,j}x_j$ such that $q(x_1,\ldots,x_n) = q'(x_1',\ldots,x_n')$. The problem is to classify quadratic forms up to this equivalence. This was solved in [Hasse, 1924] by the *Hasse–Minkowski theorem* and the *Hasse invariant*. The theorem says that two quadratic forms over a number field K are equivalent if and only if they are equivalent over all of the local fields $K_{\mathfrak{p}}$ for all primes $\mathfrak{p}$ of K. For instance for $K = \mathbf{Q}$, the rational numbers, two forms over $\mathbf{Q}$ are equivalent if and only if they are equivalent over the extensions $\mathbf{R}$, the real numbers, and the p-adic numbers $\mathbf{Q}_p$ for all prime numbers p. This reduces the problem to classification over local fields, which is handled by the Hasse invariant (apart from rank and discriminant). It is interesting to note that the definition of the Hasse invariant uses the Hilbert symbol and thus links to reciprocity from problem 9. For much more information on the theory of quadratic forms, see [O'Meara, 1971; Malyshev, 1991].

PROBLEM 12. *Extension of Kronecker's theorem on Abelian fields to any algebraic realm of rationality.*

The Kronecker–Weber theorem says that the maximal Abelian (meaning Abelian Galois group) extension $\mathbf{Q}^{ab}$ of the rational numbers is obtained by adjoining to $\mathbf{Q}$ all the roots of unity. This has two parts: on the one hand it gives an explicit construction of $\mathbf{Q}^{ab}$; on the other hand it calculates the Galois group $\mathrm{Gal}(\mathbf{Q}^{ab}/\mathbf{Q})$. The second part has been nicely generalized for any number field (and also more generally). This is the topic of class field theory, which started with [Takagi, 1920]; since then the subject has gone through several incarnations; on some of them see [Cassels and Fröhlich, 1967; Artin and Tate, 1961; Weil, 1973; Hazewinkel, 1975; Neukirch, 1986]. The first part, on explicit generation, fared less well except for the 'complex multiplication' case and local fields [Lubin and Tate, 1965]; but see also [Holzapfel, 1995].

Nowadays there is great interest in and great progress on 'non-Abelian class field theory' in the form of the conjectured Langlands correspondence. In the local case

(now proved for GL_n) this is a correspondence between representations of degree r of $\mathrm{Gal}(K^{\mathrm{sep}}/K)$, or rather a dense subgroup W_K of it called the Weil group and certain representations of $GL_r(K)$. Here K^{sep} is the separable closure of K. For the global case $GL_r(K)$ is replaced by $GL_r(\mathbf{A})$, where $\mathbf{A}$ is the ring of adèles of K. The correspondence is also supposed to satisfy a number of strong extra properties. In case $r = 1$ Abelian class field theory is recovered. No less than four invited lectures dealt with the Langlands correspondence at the latest International Congress of Mathematicians in 2002 [Li et alii, 2002]. Also there have been five *Séminaire Bourbaki* reports on the matter in recent years, giving another indication of how important the matter is considered to be [Carayol, 2000; Laumon, 2002]; and Gaitsgory and Harris in [Li et alii, 2002].

PROBLEM 13. *Impossibility of the solution of the general equation of the seventh degree by means of functions of only two variables.*

This problem is nowadays seen as a mixture of two parts: a specific algebraic (or analytic) one concerning equations of degree 7, which remains unsolved, and a '*superposition problem*': can every continuous function in n variables be written as a superposition of continuous functions of two variables? The latter problem was solved in [Arnol'd, 1957] and [Kolmogorov, 1956]: each continuous function of n variables can be written as a composite (superposition) of continuous functions of two variables.

A composite function is one obtained by substituting other functions for the variables in the first functions. So, as an example, $f(x, y, z) = F(g(x, y), h(z, k(y, z)))$ is a function of three variables that is a composite of functions of two variables. Thus, for instance, all rational functions in any number of variables, can be obtained as composites of $x + y$, $x - y$, xy and x/y.

The picture changes drastically if differentiability or analyticity conditions are imposed. For instance, there are analytic functions of n variables that cannot be written as composites of analytic functions of fewer variables.

The reason that the two parts of the problem occur together is that by Tschirnhausen transformations the general equation of degree 7 can be reduced to something of the form $X^7 + xX^3 + yX^2 + zX + 1 = 0$ (but no further), and the solutions of this equation as functions of x, y and z were considered to be candidates for functions of three variables that cannot be written as composites of functions of two variables.

PROBLEM 14. *Proof of the finiteness of certain complete systems of functions.*

The precise form of the problem is as follows: Let K be a field in between a field k and the field of rational functions $k(x_1, \ldots, x_n)$ in n variables over k: $k \subset K \subset k(x_1, \ldots, x_n)$. Is it true that $K \cap k[x_1, \ldots, x_n]$ is finitely generated over k? The motivation came from positive answers (by Hilbert for instance) in a number of important cases where there is a group, G, acting on k^n and K is the field of G-invariant rational functions. A counterexample, precisely in this setting of rings of invariants, was given by Nagata [1959]. However, in the invariants case finite generation is true if the group is reductive; this is for instance the case if G is semisimple and k is of characteristic zero [Mumford, 1965].

PROBLEM 15. *Rigorous foundation of Schubert's enumerative calculus.*

The problem is to justify and to make precise H. Schubert's '*principle of conservation of numbers*' under suitable continuous deformations. Mostly intersection numbers are involved; for instance, to prove rigorously that there are indeed 666,841,048 quadric surfaces tangent to 9 given quadric surfaces in space [Schubert, 1879]. There are a great number of such principles of conservation of numbers in intersection theory [Danilov, 1990] and cohomology and differential topology. Indeed, one version of another of this idea is often the basis of definitions in singular cases.

In spite of a great deal of progress, there remains much to be done to obtain a true *enumerative geometry* such as Schubert dreamt of. In fact, more is required than just a good intersection theory that takes care of multiplicities. One also needs to give the collection of, say, all quadric surfaces in space the structure of something like an algebraic variety, i.e. something to which intersection theory can be applied. This is a fundamental subfield of algebraic geometry, starting with the question, which goes back to Bernhard Riemann, as to on how many parameters a given kind of structure depends (how many moduli are needed in the phraseology of the 19th century, which explains the terminology 'moduli space' in algebraic geometry).

PROBLEM 16. *Problem of the topology of algebraic curves and surfaces.*

Even in its original formulation, this problem splits into two parts. The first part concerns the topology of real algebraic varieties. For instance, an algebraic real curve in the projective plane splits up in a number of ovals (topological cycles) and the question is which configurations are possible. For degree 6 this was finally solved in 1970 by D.A. Gudkov (see [Gudkov, 1992] for this and more). There are severe constraints on the configurations that are possible; early important work on this is due to Ragsdale [1906]. However, her conjectures have been fairly recently disproved by Itenberg and Viro [1996].

The second part concerns the topology of limit cycles of dynamical systems. A first problem here is the *Dulac conjecture* on the finiteness of the number of limit cycles of vector fields in the plane. For polynomial vector fields this was settled in the positive sense by Yu.S. Il'yashenko in 1970. For this and much more see [Arnol'd and Il'yashenko, 1988; Il'yashenko, 1991; Il'yashenko and Yakovenko, 1995; Roussarie, 1998].

PROBLEM 17. *Expression of definite forms by squares.*

The problem is the following. Consider a rational function of n variables over the reals which takes nonnegative values in all points where it is defined. Does it follow that it can be written as a sum of squares (of rational functions)? This was solved in [Artin, 1927], by inventing the theory of formally real fields; this subject has found other applications since. For a definite function on a real irreducible algebraic variety of dimension d the *Pfister theorem* says that no more than 2^d terms are needed to express it as a sum of squares [Pfister, 1967].

PROBLEM 18. *Building up of space from congruent polyhedra.*

In its original formulation this problem has three parts.

(18a): show that there are only finitely many types of subgroups of the group $E(n)$ of isometries of $\mathbf{R}^n$ with compact fundamental domain. This was solved in [Bieberbach, 1910]; the subgroups in question are now called *Bieberbach groups*.

(18b): the tiling of space by a single polyhedron which is not a fundamental domain as in (18a); more generally, also nonperiodic tilings of space. A *monohedral tiling* is a tiling in which all tiles are congruent to one fixed tile T. If moreover the tiling is not one that comes from a fundamental domain of a group of motions one speaks of an *anisohedral tiling*. In one sense this sub-problem was settled in [Reinhardt, 1928], who found an anisohedral tiling in $\mathbf{R}^3$; and in [Heesch, 1935], who found a non-convex anisohedral polygon in the plane that admits a periodic monohedral tiling. The tile of Heesch was actually produced as a roof tile, and such tiles form the covering of the Göttingen *Rathaus*. There also exists convex anisohedral pentagons [Kershner, 1968].

This circle of problems is still is a very lively topic today (see [Schulte, 1993] for a recent survey). For instance, the convex polytopes that can give a monohedral tiling of $\mathbf{R}^d$ have not yet been classified, even for the plane.

One important theory that emerged is that of the Penrose tilings and quasi-crystals [de Bruijn, 1997]. As another example of one of the problems that emerged, it is still unknown which polyominos tile the whole plane [Golomb, 1996]. A *polyomino* is a connected figure obtained by taking n identical unit squares and connecting them along common edges.

(18c): densest packing of spheres. This is still unsolved in general. The densest packing of circles in the plane is the familiar hexagonal one (see [Thue, 1910], and the completion of this work by Fejes Tóth [1940]). Conjecturally (indeed, the Kepler conjecture of 1610) the densest packing in three-space is the lattice packing A_3, the face-centred cubic. This packing is indeed the densest lattice packing (Gauss), but conceivably there could be denser non-lattice packings, as can happen in certain higher dimensions. In 1998 T.C. Hales and S.P. Ferguson announced a proof of the Kepler conjecture. However, only two of the eight papers involved have been published so far, both in 1997. The announced proof relies heavily on the computer checking of some 5000 special cases, a situation not dissimilar to that of 30 years ago with regard to the four-colour conjecture. Still there are grounds that the proof will turn out to be substantially correct [Oesterlé, 2000].

The Leech lattice is conjecturally the densest packing in 24 dimensions. The densest lattice packings in dimensions 1–8 are known. In dimensions 10, 11, 13 there are packings that are denser than any lattice packing [Conway and Sloane, 1988].

PROBLEM 19. *Are the solutions of the regular problems in the calculus of variations always necessarily analytic?*

This problem links to the 20th problem through the Euler–Lagrange equation of the variational calculus ([Gelfand and Fomin, 1963]; and compare §19). The variational problems meant are of the form: find a function $u : \bar{\Omega} \to \mathbf{R}$ that is of class $C^1(\Omega) \cap C^2(\Omega)$ and is such that among all functions of this class the integral $\int_\Omega F(x, u(x), p(x))\,dx$ is minimal, and such that u satisfies a Dirichlet type boundary condition $u(x) = \varphi(x)$ for $x \in \partial\Omega$. Here Ω is a bounded open set in $\mathbf{R}^n$, $\bar{\Omega}$ is its closure, $\partial\Omega$ is its boundary, and $p(x) = (\partial u/\partial x_1, \ldots, \partial u/\partial x_n)$. The function F is given and satisfies the regularity (and convexity) conditions $F \in C^2$ and $(\frac{\partial^2 F}{\partial p_i \partial p_j}) > 0$.

The corresponding Euler–Lagrange equation is

$$\sum_{i,j=1}^{n} F_{p_i p_j}(x, u, p)\frac{\partial^2 u}{\partial x_i \partial x_j} + \sum_{i=1}^{n}(F_{p_i u} p_i + F_{p_i x_i}) = F_u. \tag{1}$$

Positive results on the analyticity for nonlinear elliptic partial equations were first obtained by [Bernstein, 1904] and in more or less definite form in [Petrovskij, 1939].

PROBLEM 20. *The general problem of boundary values.*

In 1900, the general theory of boundary value problems and generalized solutions to differential equations, as Hilbert wisely specified, was very incomplete. The amount of work accomplished since is enormous in achievement and volume and includes generalized solution ideas (weak solutions) such as the *distributions* of Dirac, Sobolev and Schwartz [Vladimirov, 1989] and, rather recently for the nonlinear case, *generalized function algebras*, to avoid the difficulty that distributions do not have a good multiplication [Oberguggenberger and Rosinger, 1991; Rosinger, 1990, 1998].

PROBLEM 21. *Proof of the existence of linear differential equations having a prescribed monodromy group.*

Consider a system of n first order linear differential equations $y'(z) = A(z)y$ on the Riemann sphere $\mathbf{P}^1$ where $A(z)$ is meromorphic. Let Σ be the set of poles of $A(z)$. Such a system has an n-dimensional space S of solutions. Following a solution along a loop around one of the poles by analytic continuation gives a possibly different solution. This gives a representation of the fundamental group $\pi_1(\mathbf{P}^1 \setminus \Sigma) \to GL_n(\mathbf{C})$, the monodromy representation of the system of differential equations.

The question is now whether every representation of the fundamental group comes from a system of differential equations where it is moreover required that all the poles of $A(z)$ are simple. For a long time it was thought that this was true by the work of L. Plemelj, G. Birkhoff and I. Lappo-Danilevskij; but then in 1989 A. Bolibrukh found counterexamples. However, if extra apparent singularities are allowed: singularities where the monodromy is trivial, there is a positive solution [Beauville, 1993; Anosov and Bolibrukh, 1994].

As formulated by Hilbert the 21st problem had to do with an nth order linear differential equation of Fuchsian type $y^{(n)} + a_1 y^{(n-1)} + \cdots + a_n y = 0$ which means that a_i has at most a pole of order i. Here the answer is again negative if no apparent singularities are allowed and positive if this is allowed. In the modern literature the question is studied in the form of connections on a bundle over any Riemann surface or even in far more general situations [Röhrl, 1957; Deligne, 1970].

PROBLEM 22. *Uniformization of analytic relations by means of automorphic functions.*

This is the uniformization problem, that is, representing (most of) an algebraic or analytic manifold parametrically by single-valued functions. For instance $(\sin t, \cos t)$ and $(\frac{2u}{u^2+1}, \frac{u^2-1}{u^2+1})$, with t and u complex variables, both parametrize the Riemann surface of $z^2 + w^2 = 1$. The (complex) dimension one case was solved by H. Poincaré and P. Koebe in 1907 in the form of the *Koebe general uniformization theorem*, namely that a Riemann surface topologically equivalent to a domain in the extended complex plane is also conformally equivalent to such a domain; and the *Poincaré–Koebe theorem* or *Klein–Poincaré uniformization theorem* [Gusevskij, 1993]. For higher (complex) dimensions these questions are still largely open, as they are also for a variety of generalizations.

PROBLEM 23. *Further development of the methods of the calculus of variations.*

As in problem 19 the problem is to find curves, surfaces, ... that minimize certain integrals. Many problems in physics are formulated this way. Hilbert felt that the calculus of variations had been somewhat neglected and had a number of precise ideas of how to go further. Though there were already in 1900 a great many results in the calculus of variations [Kneser, 1900], very much more has been developed since both as regards what may be termed the classical calculus of variations [Gelfand and Fomin, 1963; Moiseev, 1993], and numerous more modern offshoots such as optimal control [Pontryagin et alii, 1962; Lions, 1968] and dynamic programming [Bellmann, 1957]. One notes also the calculus of variations in the large started in [Morse, 1934]; the theory of minimal differential geometric objects such as geodesics, minimal surfaces and Plateau's problem [Gromoll et alii, 1992; Osserman, 1969; Dao and Fomenko, 1987; Yang, 1994]; variational inequalities [Kinderlehrer and Stampacchia, 1980]; and links with convex analysis [Ekeland and Teman, 1973].

Treating variational problems as optimization problems in infinite dimensional (function) spaces brings a unifying perspective [Ioffe and Tihomirov, 1989].

3 CONCLUDING REMARKS

As is only natural, the idea of having another new stimulating list of problems for the 21st century has arisen. There was such an attempt in 1974 at the occasion of the review of the then current status of the Hilbert problems, and there are 27 groups of problems in the proceedings of that meeting [Browder, 1976]. They do not seem to have been all that successful as a guide to research. More recently, Stephen Smale formulated a list

[Smale, 1998]. Still more recently, the seven millennium problems were formulated by the new Clay Institute of Mathematics (see [Devlin, 2002] for a popular account, and go to http://www.claymath.org for the official descriptions of these seven problems: some of these are very well written indeed). Six of them are far more deeply imbedded in technically sophisticated mathematics than were the original Hilbert problems. The seven are: the Riemann hypothesis; Yang–Mills theory and the mass gap hypothesis (quantum mechanics); the P versus NP problem (mathematical programming, combinatorial optimization); the Navier–Stokes equations (fluid mechanics); the Poincaré conjecture (topology of manifolds); the Birch and Swinnerton–Dyer conjecture (arithmetic algebraic geometry); and the Hodge conjecture (algebraic geometry).

Each question carries prize money of 1 million dollars. It remains to be seen whether they will do as much as is hoped to attract brilliant young people to research mathematics. Perhaps not. For much of the 20th century there may have been a sort of general pervasive feeling that there is something like a vast, potentially complete, unique (rigid) edifice constituting mathematics. And perhaps that accounts for the feelings of (foundational) anxiety that one senses when reading accounts of the progress of mathematics on the Hilbert problems.

Today seems to be less a period of problem solving, nor a period of large theory building. Instead we seem to live in a period of discovery where new beautiful applications, interrelations and phenomena appear with astonishing frequency. It is a multiverse of many different axiom systems, of different models of even something as basic as the real numbers, of infinitely many different differential structures on the space-time, $\mathbf{R}^4$, that we live in. It is a world of many different chunks of mathematics, not necessarily provably compatible, at least until we come up with new ideas of what it means to be provable. Nor need all of mathematics be compatible. Meanwhile mathematicians go happily about the delightful business of discovering (or inventing) and describing new beauty and insights.

BIBLIOGRAPHY

Alexandrov, P.S. (ed.) 1979. *Die Hilbertschen Probleme*, Leipzig: Geest & Portig. [New ed. Frankfurt: Harri Deutsch, 1998. Original Russian ed. 1969.]

Anosov, D.V. and Bolibrukh, A.A. 1994. *The Riemann–Hilbert problem*, Braunschweig: Vieweg.

Arnol'd, V.I. 1957. 'On functions of three variables', *Dokl. Akad. Nauk, 114*, 679–681. [In Russian.]

Arnol'd, V.I. and Il'yashenko, Yu.S. 1988. 'Ordinary differential equations', in D.V. Anosov and V.I. Arnol'd (eds.), *Dynamical systems*, vol. 1, Berlin: Springer, 7–148.

Artin, E. 1927. 'Über die Zerlegung definiter Funktionen in Quadrate', *Abh. Math. Sem. Univ. Hamburg, 5*, 100–115.

Artin, E. 1928. 'Beweis des allgemeinen Reziprozitätsgesetzes', *Ibidem*, 353–363.

Artin, E. and Tate, J. 1961. *Class field theory*, Princeton: Inst. Advanced Study.

Beauville, A. 1993. 'Equations différentielles à points singuliers réguliers d'après Bolybrukh', in *Séminaire Bourbaki 1992/1993*, Soc. Math. de France, 103–120.

Bellmann, R. 1957. *Dynamic programming*, Princeton: Princeton Univ. Press.

Bernstein, S.N. 1904. 'Sur la nature analytique des solutions des équations aux dérivées parteilles des second ordre', *Math. Ann., 59*, 20–76.

Bieberbach, L. 1910. 'Über die Bewegungsgruppen des n-dimensionalen euklidisches Raumes mit einem endlichen Fundamentalbereich', *Göttingen Nachr.*, 75–84.

Blaschke, W. 1936. 'Integralgeometrie II', *Abh. Math. Sem. Hamburg, 11*, 359–366.
Bricard, R. 1896. 'Sur une question de géométrie relative aux polyèdres', *Nouv. ann. math., 15*, 331–334.
Browder, F.E. (ed.) 1976. *Mathematical developments arising from Hilbert's problems*, Providence: Amer. Math. Soc.
Busemann, H. 1961. 'Geometries in which the planes minimize area', *Ann. mat. pura appl., 55*, 171–190.
Carayol, H. 2000. 'Preuve de la conjecture de Langlands locale pour GL_n: Travaux de Harris-Taylor et Henniart', in *Séminaire Bourbaki 1998/1999*, Exposé 857, Soc. Math. de France, 191–244.
Cassels, J.W.S. and Fröhlich, A. (eds.) 1967. *Algebraic number theory*, San Diego: Academic Press.
Cohen, P.J. 1963, 1964. 'The independence of the continuum hypothesis', *Proc. Nat. Acad. Sci. USA, 50*, 1143–1148; *51*, 105–110.
Conrey, J.B. 1989. 'More than two fifths of the zeroes of the Riemann zetafunction are on the critical line', *J. reine und angew. Math. 399*, 1–26.
Conway, J.H. and Sloane, N.J.A. 1988. *Sphere packings, lattices and groups*, Berlin: Springer.
Danilov, V.I. 1990. 'Intersection theory on an algebraic variety', in M. Hazewinkel (ed.), *Encyclopaedia of mathematics*, vol. 5, Dordrecht: Kluwer, 153–155.
Dao, Ch.T, and Fomenko, Ch.T. 1987. *Minimal surfaces and the Plateau problem*, Moscow: Nauka. [In Russian.]
de Bruijn, N.G. 1997. 'Penrose tiling', in M. Hazewinkel (ed.), *Encyclopaedia of mathematics*, vol. 11, Dordrecht: Kluwer, 404–405.
Dehn, M. 1901. 'Über den Rauminhalt', *Math. Ann., 55*, 465–478.
Deligne, P. 1970. *Equations différentielles à points singuliers réguliers*, Berlin: Springer.
Deligne, P. 1974, 1980. 'La conjecture de Weil', *Publ. Math. IHES, 43*, 273–308; *52*, 137–252.
Devlin, K. 2002. *The millenium problems*, New York: Basic Books.
Ekeland, I. and Teman, R. 1973. *Analyse convexe et problèmes variationels*, Paris: Dunod and Gauthier–Villars.
Fejes Tóth, L. 1940. 'Über einem geometrischen Satz', *Math. Ztsch., 46*, 79–83.
Fel'dman, N.I. and Nestorenko, Yu.V. 1998. *Transcendental numbers*, Berlin: Springer.
Gelfand, I.M. and Fomin, S.V. 1963. *Calculus of variations*, Englewood Cliffs, NJ: Prentice Hall.
Gleason, A.M. 1952. 'Groups without small subgroups', *Ann. of math., 56*, 193–212.
Gödel, K. 1931. 'Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I', *Monatshefte Math. Phys., 38*, 173–198. [See §71.]
Gödel, K. 1939. 'Consistency proof for the generalized continuum hypothesis', *Proc. Nat. Acad. Sci. USA, 25*, 220–224.
Golomb, S.W. 1996. 'Tiling rectangles with polyominoes', *Math. intelligencer, 18*, no. 2, 38–47.
Gray, J.J. 2000. *The Hilbert challenge*, Oxford: Oxford University Press.
Gromoll, D., Klingenberg, W. and Meyer, W. 1992. *Riemansche Geometrie im grossen*, Berlin: Springer.
Gudkov, D.A. 1992. 'Real algebraic variety', in M. Hazewinkel (ed.), *Encyclopaedia of mathematics*, vol. 8, Dordrecht: Kluwer, 2–4.
Gusevskij, N.A. 1993. 'Uniformization', in M. Hazewinkel (ed.), *Encyclopaedia of mathematics*, vol. 9, Dordrecht: Kluwer, 321–323.
Hamel, G. 1903. 'Über die Geometrieen in denen die Graden die kürzesten sind', *Math. Annalen, 57*, 231–264.
Hasse, H. 1924. 'Äquivalenz quadratischer formeln in einem beliebigen algebraischer Zahlkörper', *J. reine und angew. Math., 153*, 113–130.
Hazewinkel, M. 1975. 'Local class field theory is easy', *Adv. math., 18*, 148–181.

Hazewinkel, M. 2000. 'Hilbert problems', in M. Hazewinkel (ed.), *Encyclopaedia of mathematics*, vol. 12, Dordrecht: Kluwer, 257–261.

Heesch, H. 1935. 'Aufbau der Ebene aus kongruenten Bereiche', *Nachr. Ges. Wiss. Göttingen, new ser., 1*, 115–117.

Hirschfield, J. 1990. 'The nonstandard treatment of Hilbert's fifth problem', *Trans. Amer. Math. Soc.*, *321*, 379–400.

Holzapfel, R.-P. 1995. *The ball and some Hilbert problems*, Basel: Birkhäuser.

Il'yashenko, Yu. 1991. *Finiteness theorems for limit cycles*, Providence: Amer. Math. Soc.

Il'yashenko, Yu. and Yakovenko, S. (eds.) 1995. *Concerning the Hilbert 16th problem*, Providence: Amer. Math. Soc.

Ioffe, A.D. and Tihomirov, V.M. 1989. *Theory of extremal problems*, Amsterdam: North-Holland.

Itenberg, I. and Viro, O. 1996. 'Patchworking algebraic curves disproves the Ragsdale conjecture', *Math. intelligencer*, *18*, no. 4, 19–28.

Kantor, J.-M. 1996. 'Hilbert's problems and their sequels', *Math. intelligencer*, *18*, no. 1, 21–34.

Kershner, R.B. 1968. 'On paving the plane', *Amer. math. monthly*, *75*, 839–844.

Kinderlehrer, D. and Stampacchia, G. 1980. *An introduction to variational inequalities and their applications*, San Diego: Academic Press.

Kneser, A. 1900. *Lehrbuch der Variationsrechnung*, Braunschweig: Vieweg.

Kolmogorov, A.N. 1933. *Grundbegriffe der Wahrscheinlichkeitsrechnung*, Berlin: Springer. [See §75.]

Kolmogorov, A.N. 1956. 'On the representation of continuous functions of several variables as superpositions of continuous functions of fewer variables', *Dokl. Akad. Nauk SSSR, 108*, no. 2, 179–182. [In Russian.]

Laumon, G. 2002. 'La correspondence de Langlands sur les corps de fonctions (d'après Laurent Lafforgue)', in *Séminaire Bourbaki 1999/2000*, Exposé 873, Soc. Math. de France, 2002, 207–265.

Lawrence, R.J. 1996. 'An introduction to topological field theory', in L.H. Kaufmann (ed.), *The interface of knots and physics*, Providence: Amer. Math. Soc., 89–128.

Levinson, N. 1974. 'More than one third of the zeros of the Riemann zeta-function are on $\mathrm{Re}(s) = 1/2$', *Adv. math.*, *13*, 383–436.

Li, Ta Tsien et alii (eds.) 2002. *Proceedings of the international congress of mathematicians, ICM 2002* Beijing: Higher Education Press.

Lions, J.L. 1968. *Contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles*, Paris: Dunod and Gauthier–Villars.

Lubin, J, and Tate, J. 1965. 'Formal complex multiplication in local fields', *Ann. of math.*, *81*, 380–387.

Malyshev, A.V. 1991. 'Quadratic form', in M. Hazewinkel (ed.), *Encyclopaedia of mathematics*, vol. 7, Dordrecht: Kluwer, 378–382.

Moiseev, N.N. 1993. 'Variational calculus', in M. Hazewinkel (ed.), *Encyclopaedia of mathematics*, vol. 9, Dordrecht: Kluwer, 1993, 372–377.

Montgomery, D. and Zippin, L. 1952. 'Small subgroups of finite dimensional groups', *Ann. of math.*, *56*, 213–241.

Morse, M. 1934. *The calculus of variations in the large*, Providence: Amer. Math. Soc.

Mumford, D. 1965. *Geometric invariant theory*, Berlin: Springer.

Nagata, M. 1959. 'On the fourteenth problem of Hilbert', *Amer. j. math.*, *81*, 766–772.

Nagel, E. and Newman, J.R. 1959. *Gödel's proof*, London: Routledge & Kegan Paul.

Neukirch, J. 1986. *Class field theory*, Berlin: Springer.

O'Meara, O.T. 1971. *Introduction to quadratic forms*, 2nd ed., Berlin: Springer.

Oberguggenberger, M. and Rosinger, E.E. 1991. *Solutions of continuous nonlinear PDE's through order completion*, pt. 1, Pretoria: University of Pretoria.
Oesterlé, J. 2000. 'Densité maximale des empilements de sphères en dimension 3', in *Séminaire Bourbaki 1998/1999*, Exposé 863, Soc. Math. de France, 405–413.
Osserman, R. 1969. *A survey of minimal surfaces*, New York: van Nostrand Reinhold. [Repr. New York: Dover, 1986.]
Parshin, A.N. 1993. 'Zeta function', in M. Hazewinkel (ed.), *Encyclopaedia of mathematics*, vol. 9, Dordrecht: Kluwer, 527–534.
Petrovskij, I.G. 1939. 'Sur l'analyticité des solutions des systèmes d'équations différentielles', *Mat. sbornik*, *5*, 3–70.
Pfister, A. 1967. 'Zur Darstellung definiter Funktionen als Summe von Quadraten', *Inv. math.*, *4*, 229–237.
Pheidas, T. 1994. 'Extensions of Hilbert's tenth problem', *J. symbolic logic*, *59*, 372–397.
Pogorelov, A.V. 1973. 'A complete solution of Hilbert's fourth problem', *Sov. math. doklady*, *14*, 46–49.
Pogorelov, A.V. 1979. *Hilbert's fourth problem*, New York: Winston & Wiley.
Pontryagin, L.S., Boltyantskij, V.G., Gamkrelidze, R.V. and Mishchenko, E.F. 1962. *Mathematical theory of optimal processes*, New York: Wiley.
Putnam, H. 1960. 'An unsolvable problem in number theory', *J. symbolic logic*, *25*, 220–232.
Ragsdale, V. 1906. 'On the arrangement of the real branches of plane algebraic curves', *Amer. j. math.*, *28*, 377–404.
Reid, C. 1970. *Hilbert*, New York: Springer.
Reinhardt, K. 1928. 'Zur Zerlegung Euklidische Räume in kongruente Polytope', *Sitz.-Ber. Preuss. Akad. Wiss.*, 150–155.
Röhrl, H. 1957. 'Das Riemann-Hilbertsche Problem der Theorie der linearen Differentialgleichungen', *Math. Annalen*, *133*, 1–25.
Rosinger, E.E. 1990. *Non-linear partial differential equations*, Amsterdam: North-Holland.
Rosinger, E.E. 1998. *Parametric Lie group actions on global generalised solutions of nonlinear PDEs*, Dordrecht: Kluwer.
Roussarie, R. 1998. *Bifurcation of planar vector fields and Hilbert's sixteenth problem*, Basel: Birkhäuser.
Sah, C.-H. 1979. *Hilbert's third problem: scissors congruence*, London: Pitman.
Sawin, S. 1996. 'Links, quantum groups, and TQFTs', *Bull. Amer. Math. Soc.*, *33*, 413–445.
Schubert, H.C.H. 1879. *Kalkül der abzählenden Geometrie*, Leipzig: Teubner.
Schulte, E. 1993. 'Tilings', in P.M. Gruber and J.M. Wills (eds.), *Handbook of convex geometry*, vol. B, Amsterdam: North-Holland, 899–932.
Segal, G. 1988. 'The definition of conformal field theory', in K. Bleuler and M. Werner (eds.), *Differential geometrical methods in theoretical physics*, Dordrecht: Kluwer, 165–171.
Segal, G. 1991. 'Geometric aspects of quantum field theory', in *Proceedings ICM Kyoto 1990*, vol. 2, Berlin: Springer, 1387–1396.
Selberg, A. 1942. 'On the zeros of the zeta-function of Riemann', *Der Kong. Norske Vidensk. Selsk. Forhand.*, *15*, 59–62.
Shafarevich, I.R. 1950. 'General reciprocity laws', *Mat. sbornik*, *26*, 13–146. [In Russian.]
Smale, S. 1998. 'Mathematical problems for the next century', *Math. intelligencer*, *20*, no. 2, 7–15.
Stepanov, S.A. 1992. 'Reciprocity laws', in M. Hazewinkel (ed.), *Encyclopaedia of mathematics*, vol. 8, 10–11.
Szabo, Z.I. 1986. 'Hilbert's fourth problem', *Adv. math.*, *59*, 185–301.
Takagi, T. 1920. 'Über eine Theorie des relativ-abelschen Zahlkörpers', *J. Coll. Sci. Imp. Univ. Tokyo*, *41*, 1–132.

Thue, A. 1910. ‘Über die dichteste Zuzammenstellung von kongruenten Kreisen in einer Ebene’, *Skr. Vidensk-Selsk Christ.*, *1*, 1–9.

Turaev, V.G. 1994. *Quantum invariants of knots and 3-manifolds*, Berlin: de Gruyter.

van de Lune, J., te Riele, J.J., and Winter, D. 1986. ‘On the zeros of the Riemann zeta-function in the critical strip IV’, *Math. of comp.*, *46*, 667–681.

Vladimirov, V.S. 1989. ‘Generalized functions, Space of generalized functions, Product of generalized functions, Derivative of generalized functions’, in M. Hazewinkel (ed.), *Encyclopaedia of mathematics*, vol. 4, Dordrecht: Kluwer, 1989, 228–235.

von Neumann, J. 1932, *Mathematische Grundlagen der Quantummechanik*, Berlin: Springer. [See §69.]

Weil, A. 1973. *Basic number theory*, Berlin: Springer.

Witten, E. 1988. ‘Topological quantum field theory’, *Comm. math. physics*, *117*, 353–386.

Yandell, B.H. 2002. *The honors class. Hilbert’s problems and their solvers*, Natick, MA: Peters.

Yang, K. 1994. *Complete minimal surfaces of finite total curvature*, Dordrecht: Kluwer.

CHAPTER 58

LORD KELVIN, BALTIMORE LECTURES ON MATHEMATICAL PHYSICS ((1884), 1904)

Ole Knudsen

Kelvin gave a comprehensive, if somewhat idiosyncratic, survey of 19th-century work on mathematical elasticity theory as applied to the luminiferous ether and the wave theory of light.

First publication. William Thomson, *Notes of lectures on molecular dynamics and the wave theory of light*, Baltimore: Johns Hopkins University, 1884. 328 pages + index. [A hand-written, verbatim report of the lectures, reproduced by the so-called 'papyrograph' process and sent to the members of the audience; printed version in [Kargon and Achinstein, 1987, 7–263].]

Second publication. Lord Kelvin, *Baltimore lectures on molecular dynamics and the wave theory of light. Founded on Mr A.S. Hathaway's stenographic report of twenty lectures delivered in Johns Hopkins University, Baltimore, in October, 1884: followed by twelve appendices on allied subjects*, London: C.J. Clay and Sons, 1904. xxii + 694 pages.

German translation. Vorlesungen über Moleculardynamik und die Theorie des Lichtes (trans. B. Weinstein), Leipzig and Berlin: Teubner, 1909.

Related articles: Green (§30), Thomson and Tait (§40), Maxwell (§44), Lorentz (§60), Einstein (§63).

1 BIOGRAPHY

William Thomson, Lord Kelvin (1824–1907) was educated at the Universities of Glasgow and Cambridge, where he graduated as second wrangler and first Smith's prizeman in 1845 (see also §40.1). After spending some months in Paris working in Victor Regnault's laboratory on the thermal properties of steam, he became, in 1846, professor of natural philosophy in Glasgow, a position he held until his retirement in 1899. He is known today for his formulation of the two laws of thermodynamics and his invention of an absolute

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

scale of temperature (the Kelvin scale); in his own time he was famous for his work on the Atlantic telegraph for which he received a knighthood in 1866. He was raised to the peerage in 1892, choosing the name 'Kelvin' from a river that runs through the Glasgow University campus; in the following this name will be used throughout.

2 MATHEMATICAL FIELD THEORY

Kelvin's first important work was in the mathematical theory of electricity and magnetism [Knudsen, 1985; Smith and Wise, 1989, chs. 7 and 8]. His main inspiration came from J.B.J. Fourier and Michael Faraday; he also admired the work of George Green (§30). From [Fourier, 1822], which he read as a 16-year old undergraduate, he saw how a highly developed mathematical theory of heat conduction could be created when based solely on macroscopic, observable quantities like temperature and quantity of heat, without recourse to dubious hypotheses on the nature of heat and molecular action (§26). In [Faraday, 1839–1855] he was impressed by the attempt to replace action-at-a-distance theory of electricity and magnetism by the concept of contiguous action, the embryo of the field concept (vol. 1, 360-364, 380). This dual inspiration led to Kelvin's discovery in 1842 [Thomson, 1872, 1–14] of the analogy between Fourier's mathematical description of the temperature distribution in a steady heat flow and electrostatic potential (both satisfy the Laplace equation) and likewise between heat flux and electric force (both are proportional to the gradient of, respectively, temperature and potential).

The analogies allowed Thomson to translate results that are almost self-evident in heat conduction into not so evident propositions in electrostatics, and vice versa. The analogies also lay behind his invention of the method of electric images that furnished a synthetic, geometrical method of constructing solutions for potential and electric force in a system of conducting surfaces and point charges [Thomson, 1872, 144–146]. Finally, the heat analogy played a fundamental role in making Faraday's non-mathematical field-theoretical concepts mathematically respectable. Faraday had thought that his discovery of dielectric effects and his description of these as caused by local actions propagated through the dielectric material would conflict with the mathematical electrostatics of C. Coulomb and S.D. Poisson as based on action at a distance. In 1845 Kelvin argued that the mathematical equivalence between electrostatics and heat conduction showed that there was no a priori reason why opposing physical views—action at a distance or action propagated through a field—could not lead to identical mathematical theories. He went on to show, drawing on Poisson's theory of magnetic polarisation, that the heat analogy could be extended to dielectric actions, media with different dielectric constants being represented by media of different conductivities for heat [Thomson, 1872, 26–37].

Kelvin continued to work on mathematical field theory until about 1856 when he became absorbed in the Atlantic telegraph cable enterprise. He published in 1848 a formulation of the Dirichlet principle and applied it afterwards to electrostatics and hydrodynamics where he showed that the function that it states to be a minimum is the energy of the system ([Thomson, 1872, 139–143]; see [Cross, 1985, 140–141] and [Knudsen, 1985, 159–161]). The earliest formulation of the Stokes integral theorem is found in a letter of 2 July 1850, from Kelvin to G.G. Stokes [Cross, 1985, 143–144]; Kelvin had collaborated with Stokes

on a series of notes on hydrodynamics and corresponded frequently with him till Stokes died in 1903 [Wilson, 1990]. He also worked with Stokes on the mathematical theory of magnetism in which he drew on the mathematical analogy between the magnetic field and the velocity field of an incompressible fluid. He found the two different representations corresponding to the modern vectors **B** and **H** which satisfied respectively the conditions $\operatorname{div}\mathbf{B} = 0$ and $\operatorname{curl}\mathbf{H} = 0$, the 'solenoidal' and 'lamellar' distributions, as he called them because the first described the magnetic field from a distribution of Ampèrean currents and the second that from a distribution of magnetized sheets [Smith and Wise, 1989, 263–275; Knudsen, 1985, 161–164].

3 THE ETHER AS AN ELASTIC SOLID

A.J. Fresnel's theory of light as transverse waves had caused mathematicians like A.L. Cauchy and Green to develop a mathematical theory of stress and strain in an elastic solid and to attempt with some, but not total, success to make the theory comprise optical effects such as reflection, refraction, and double refraction. Kelvin's friend Stokes took part in this attempt: in 1845 he published a paper on the mechanics of fluids and solids [Stokes *Papers*, vol. 1, 75–129] containing a Section entitled 'Reflections on the constitution, and equations of motion of the luminiferous ether in vacuum' [*ibidem*, 124–129]. Stokes's elasticity theory inspired Kelvin to a paper entitled 'On a mechanical representation of electric, magnetic, and galvanic [electromagnetic] forces', published in 1847 [Thomson *Papers*, vol. 1, 76–80], where he described a new mathematical analogy in which the electric force from a point charge was represented by the elastic displacement in an elastic solid in equilibrium under the action of straining forces on its surface, while the magnetic force from a small magnetic dipole, and the electromagnetic force from an electric current encircling an infinitesimal area, both were modeled by the *curl* of the displacement, that is, by the differential rotation of a volume element of the solid. This opened a vision of a grand theory in which not only optics, but the whole of electricity and magnetism as well, would be subsumed under a mechanical theory of the ether as an elastic solid, a vision that would haunt Kelvin for the rest of his life; as he said in a letter to G.F. FitzGerald in 1896: 'I have not had a moment's peace or happiness in respect to electro-magnetic theory since Nov. 28, 1846. [. . .] All this time I have been liable to fits of ether dipsomania, kept away at intervals only by rigorous abstention from thought on the subject' [Thompson, 1910, vol. 2, 1065].

4 THE PHYSICAL FOUNDATION OF THE *BALTIMORE LECTURES*

From the mid 1850s Kelvin's work on the Atlantic telegraph and his other commercial engagements, as well as his collaboration with P.G. Tait on their joint *Treatise on natural philosophy* (1867: see §40), kept the dipsomania at bay for more than 15 years. When preparing a *Reprint* of earlier papers [Thomson, 1872], he wrote a great number of footnotes and additional Sections, which brought on a few sporadic attacks; but it was his lectures at Baltimore in 1884 and his subsequent 20 years' work of revision of them for the 1904 edition that turned his occupation with the ether into something like an obsession.

The physical basis of the original lectures as well as the 1904 revision consisted in an attempt to use continuum mechanics to describe transverse waves in an elastic solid ether and to couple these with the internal vibrations of molecules embedded in the ether and having discrete mechanical degrees of freedom, in order to see how far such a theory could reproduce quantitatively known optical effects. Already in 1884 this approach was beginning to be outdated. Maxwell's electromagnetic theory of light was, at least in principle, accepted by a number of physicists and became so much more universally after Heinrich Hertz's experiments on electromagnetic waves in 1889. It did not preclude mechanical theories of the ether—a number of such attempts were made by younger disciples of Maxwell [Hunt, 1987; Buchwald, 1985; Darrigol, 2000, ch. 5]—but it made attempts to understand light as a purely mechanical phenomenon without a similar mechanical understanding of electromagnetic fields seem outdated. By 1904 Kelvin's views on the ether and on Maxwell's theory were completely obsolete. His stubborn refusal to accept Maxwell's concept of displacement current and his insistence that electrostatic action had to be propagated as condensational waves in the ether had left him hopelessly behind the newer trend in physics where dynamical explanations or models had been replaced by consistent mathematical structures that could reproduce known relations of physical systems [Knudsen, 1985; Wise and Smith, 1987]. In his 1904 review of the *Baltimore lectures* J. Larmor said that '30 years ago' Kelvin's work 'would probably have been received with universal acclaim', but that 'most of us are now wedded to the electric theory of light [. . .] which forms a consistent scheme of the relations of electricity and radiation' [Wise and Smith, 1987, 323].

5 THE CONTENTS OF THE *BALTIMORE LECTURES*

As is indicated in the title above of the 1904 edition, the lectures were delivered in Johns Hopkins University, Baltimore, in October 1884: the contents of that edition are summarised in Table 1. They take the form of a treatise on the mathematical theory of elasticity in Sections distinguished by the caption 'molar', and intertwined with Sections, labeled 'molecular', which treat the mechanical vibrations of systems having a denumerable number of degrees of freedom.

Everywhere the strict mathematical treatment is interrupted by examples, applications of the abstract theory to optics in particular, accounts of experiments, remarks on Kelvin's high opinion of Green, Stokes and Lord Rayleigh, his dislike of the style of J.L. Lagrange (§16) and of Poisson, his arguments against Maxwell's theory, and other asides occurring to him during the lectures. Dated footnotes and insertions between square brackets bear witness to insights obtained by Kelvin during the period of revision. Out of this tangled mass of material it is, however, possible, from the 'molar' Sections of Lectures II–IV, to distil a consistent elasticity theory. In the following discussion a modernized notation will be used.

Kelvin begins by postulating the potential energy per unit volume, E, of a strained elastic solid as a quadratic function of the six elements of the symmetric strain tensor:

$$e_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right), \tag{1}$$

Table 1. Contents by Lectures of the 1904 edition. This table follows the original list on pp. ix–xxi, but with many reductions. In the first column the lectures take roman numerals, then the appendices take capital letters.

	Page	**Topics**
I	5	Introductory: Wave theory of light. Ordinary and anomalous dispersion. 'Electromagnetic theory of light'. Direction of vibrations. Refraction and reflection. Double refraction.
II	22	Molar: Dynamics of elastic solid. General equations of motion. Molecular: Dynamics of a row of connected particles.
III	34	Molar: Dynamics of elastic solid. Molecular: Variations of complex serial molecule.
IV	41	Molar: Equations of motion of elastic solid. 'Electromagnetic theory of light' wants dynamical foundation. Condensational waves travelling outwards from a point.
V	52	Molar: Vibrations of air round a tuning fork. Molecular: Vibrations of serial molecule. Fluorescence.
VI	61	Molar: Ratio of rigidity to compressibility. Velocity of groups of waves through transparent substances. Molecular: Vibrations of serial molecule.
VII	71	Molecular: Vibrations of serial molecule. Metallic reflection. Double refraction.
VIII	80	Molar: Solutions for distortional waves. The blue of the sky. Molecular: Fluorescence and phosphorescence; refraction; anomalous dispersion.
IX	94	Molar: Interference. Loss of energy in waves. Dynamics of absorption, of anomalous dispersion. Molecular: seven vibrating particles. Refraction; refractive index.
X	108	Molar: Energy of waves. Fourier's theorem. Deep-sea waves. Molecular: Polarization by reflection, double refraction. Anomalous dispersion, fluorescence, phosphorescence, and radiant heat, discoverable by dynamics alone.
XI	122	Molar: Aelotropy. Green's 21 moduli of elasticity. His full theory. Most general plane wave.
XII	135	Molar: Three sets of plane waves; wave-surface with three sheets. Energy of condensational waves in ether. Molecular: Mutual force between atom and ether. Dispersion; refractivity. Critical periods. Continuity in undulatory theory.
XIII	163	Molecular: Vibrator of seven periods. Dynamical wave machine. Molar: Aelotropy resumed. Dynamics for wave surface. Molecular: Sellmeier's dynamical theory of dark lines. Photographs of anomalous dispersion by Henri Becquerel.
XIV	185	Molecular: Motion of ether with embedded molecules. Molar: Mathematical investigation of spherical waves in elastic solid.

Table 1. (*Continued*)

	Page	Topics
XV	220	Molecular: Excitation of synchronous vibrators in molecule by light. Molar: Compressibility. Double refraction; stress theory.
XVI	260	Molar: Mechanical value of sunlight; density, rigidity of ether. Velocities, number, and masses of stars.
XVII	279	Molecular: Molecular dimensions. Kinetic theory of gases. Dynamics of the blue sky. Vibrations of polarized light are perpendicular to plane of polarization.
XVIII	324	Molar: Reflection of light. Fresnel's laws, Green's theory. Opacity and reflectivity of metals.
XIX	408	Molecular: 7 mutually interacting particles, numerical solution. Molar: Navier–Poisson doctrine disproved. Molecular: Interactions between atoms and ether. Molar: Adamantinism. Double refraction.
XX	436	Molecular: Chiral rotation of plane of polarization. Molar: Chiral inertia in wave-motion. Magneto-optic rotation. Molecular: Electro-etherial theory of velocity of light.
A	468	On the motion produced in an infinite elastic solid by the motion through the space occupied by it of a body acting on it only by attraction or repulsion.
B	486	19th-century clouds over the dynamical theory of heat and light.
C	528	On the disturbance produced by two particular forms of initial displacement in an infinitely long material system for which the velocity of periodic waves depends on the wave-length.
D	532	On the clustering of gravitational matter in any part of the universe.
E	541	Aepinus atomized.
F	569	Dynamical illustrations of the magnetic and the helicoidal rotatory effects of transparent bodies on polarized light.
G	584	Hydrokinetic solutions and observations.
H	602	On the molecular tactics of a crystal.
I	643	On the elasticity of a crystal according to Boscovich.
J	662	Molecular dynamics of a crystal.
K	681	On variational electric and magnetic screening.
L	688	Electric waves and vibrations in a submarine telegraph wire. [End 694.]

where $\mathbf{u} = (u_1, u_2, u_3)$ is the displacement from equilibrium of a volume element at the point $\mathbf{r} = (x_1, x_2, x_3)$. That is,

$$E = \frac{1}{2} \sum_{i,j=1}^{6} c_{ij} e_i e_j. \tag{2}$$

Here the e's have been renumbered so that $e_{11} = e_1$ etc. and $2e_{23} = e_4$, $2e_{31} = e_5$, $2e_{12} = e_6$. If we renumber the six elements t_{ij} of the stress tensor $\mathbf{T}$ in the same way,

then we have

$$t_i = \frac{\partial E}{\partial e_i} = \sum_{j=1}^{6} c_{ij} e_j. \tag{3}$$

In principle there are 36 elastic constants c_{ij}, but since they are symmetric their number reduces to 21 in the most general case and may be reduced even further if the properties of the elastic body have some kind of symmetry. In the simplest case of a homogeneous and isotropic elastic solid (such as the free ether must be supposed to be) there are only two independent elastic constants, which may be chosen in various ways. Kelvin uses two different choices that he denotes respectively A, B and m, n; here we will choose the more familiar so-called Lamé constants λ, μ. In this case (2) and (3) become respectively:

$$E = \frac{1}{2}\lambda\left(\sum_{i=1}^{3} e_{ii}\right)^2 + \mu \sum_{i,j=1}^{3} e_{ij}^2 \tag{4}$$

and

$$t_{ij} = \frac{\partial E}{\partial e_{ij}} = \lambda\vartheta\delta_{ij} + 2\mu e_{ij}, \quad \text{where } \vartheta = \sum_{i=1}^{3} e_{ii} \text{ and } \delta_{ij} \text{ is the Kronecker symbol.} \tag{5}$$

The general equations of motion for an elastic solid of mass density ρ are, in the absence of external forces,

$$\rho \frac{\partial^2 u_i}{\partial t^2} = \sum_{j=1}^{3} \frac{\partial t_{ij}}{\partial x_j}. \tag{6}$$

In the homogeneous and isotropic case these become, by insertion of (5),

$$\rho \frac{\partial^2 \mathbf{u}}{\partial t^2} = (\lambda + \mu)\nabla(\nabla \cdot \mathbf{u}) + \mu \nabla^2 \mathbf{u}. \tag{7}$$

The general solution to this equation can be split in two different waves such that $\mathbf{u} = \mathbf{u}_1 + \mathbf{u}_2$, where

$$\rho \frac{\partial^2 \mathbf{u}_1}{\partial t^2} = \mu \nabla^2 \mathbf{u}_1, \quad \text{where } \nabla \cdot \mathbf{u}_1 = 0, \tag{8}$$

and

$$\rho \frac{\partial^2}{\partial t^2}(\nabla \cdot \mathbf{u}_2) = (\lambda + 2\mu)\nabla^2(\nabla \cdot \mathbf{u}_2), \quad \text{where } \nabla \times \mathbf{u}_2 = \mathbf{0}. \tag{9}$$

This split can, as Kelvin proves, be made in one and only one way. The first wave, $\mathbf{u}_1$, is a transversal (Kelvin calls it 'distortional') wave with velocity $\sqrt{(\mu/\rho)}$, while $\mathbf{u}_2$ is a longitudinal (Kelvin says 'condensational') wave with a different velocity of $\sqrt{[(\lambda + 2\mu)/\rho]}$. Although it is the transversal waves that are relevant for the theory of light, Kelvin

devotes some time in Lectures IV to VI to the study of longitudinal waves exemplified with sound waves round a tuning fork.

Kelvin then treats the theory of light in depth from Lecture VIII to the end. Using the mathematical theory of transversal waves outlined above, he often mingles these purely 'molar' results with results from the purely 'molecular' Sections, reflecting his belief that optical effects can only be explained satisfactorily by coupling transversal ether waves with the vibrations of material molecules. Already in his introductory Lecture I he presents a 'rude mechanical model' of ponderable matter consisting of rigid spherical shells within each other, connected by springs, and with a massive central nucleus connected by springs to the innermost shell (pp. 12–13). In Lecture II he introduces a different mechanical model called a 'serial molecule', consisting of a linear arrangement of particles connected by springs and he investigates the dynamics of this system, which is more amenable to exact calculations than the more realistic spherical molecule, throughout the succeeding lectures. Molecules of one or the other kind embedded in the ether, having requisite frequences of internal vibrations, will explain dispersion as well as anomalous dispersion (Lecture IX, X, XII, XIII) and may also be applied to such phenomena as fluorescence and phosphorescence (Lecture IX), and optical rotation (Lecture XX). In Lecture XX, which was rewritten in 1903 after the electron had been universally accepted as a universal constituent of matter, the molecular vibrators become 'electrionic' vibrators, 'electrion' being Kelvin's preferred word for what everybody else denoted as 'electron'. Kelvin's theory of 'electrions' is more fully explained in Appendices A and E.

As Table 1 shows, the appendices consist of a motley of reprints of earlier writings; some of them (Appendices F and G) date back to the 1850s and 1860s, but most contain fairly recent work. Appendix B is Kelvin's famous Friday Evening Lecture to the Royal Institution in 1900 on the two clouds obscuring the 'beauty and clearness of the dynamical theory, which asserts heat and light to be modes of motion'. The first cloud is caused by the Michelson-Morley experiment of 1887 showing that the 'ether wind' required by the theory of aberration of light is non-existent. Although the contraction suggested by G.F. FitzGerald and H.A. Lorentz seems to furnish an escape from this conclusion, 'we must still regard Cloud No. I. as very dense'. The second cloud comes from the Boltzmann–Maxwell doctrine in the kinetic theory of gases according to which the internal energy in a gas will be distributed equally between all degrees of freedom in the gas. This so-called 'equipartition theorem' leads to a relation between the ratio of the specific heat at constant pressure to that at constant volume and the number of degrees of freedom. This relation is verified experimentally for four monatomic gases (mercury vapour, argon, helium, and krypton) if these are considered as having three degrees of freedom per atom, but fails hopelessly if one notes that each of them has a large number of spectral lines and so must have a correspondingly large number of vibrational degrees of freedom per atom. Kelvin quotes Lord Rayleigh as saying that what is wanted 'is some escape from the destructive simplicity of the general conclusion' and he ends his talk by stating that 'the simplest way of arriving at this desired result is to deny the conclusion'. Within a few years Kelvin's two clouds were to be dispelled by, respectively, Albert Einstein's relativity theory of 1905 (compare §63) and the same man's quantum theory of specific heats of 1906.

6 CONCLUSION

In the history of physics Kelvin's *Baltimore lectures* has been seen, ever since its appearance, as a desperate rearguard fight of the last prominent adherent of an uncompromisingly mechanical world view against newer trends in theoretical physics such as Henri Poincaré's conventionalism or the electromagnetic world view that had become increasingly fashionable round the turn of the century [Darrigol, 2000, ch. 9]. Even in Britain, where Kelvin enjoyed an enormous respect, the Lectures met with little acclaim, as Larmor's review quoted above witnesses. The general reaction seems to have been one of polite indifference, and it is hard to point to a physicist whose work was influenced by the book. Yet for the historian it is of interest as a monument to the many attempts earlier in the 19th century to create a mechanical theory of light and matter and their interactions, and to the successes and failures of these attempts.

BIBLIOGRAPHY

Buchwald, J.Z. 1985. *From Maxwell to microphysics: aspects of electromagnetic theory in the last quarter of the nineteenth century*, Chicago: Chicago University Press.

Cross, J.J. 1985. 'Integral theorems in Cambridge mathematical physics, 1830–55', in [Harman 1985], 112–148.

Darrigol, O. 2000. *Electrodynamics from Ampère to Einstein*, Oxford: Oxford University Press.

Faraday, M. 1839–1855. *Experimental researches in electricity*, 3 vols., London: Taylor & Francis. [Repr. New York: Dover, 1965.]

Fourier, J.B.J. 1822. *Théorie analytique de la chaleur*, Paris: Firmin Didot. [See §26.]

Harman, P.M. (ed.) 1985. *Wranglers and physicists: studies on Cambridge physics in the nineteenth century*, Manchester: Manchester University Press.

Hunt, B.J. 1987. "'How my model was right": G.F. Fitzgerald and the reform of Maxwell's theory', in (Kargon and Achinstein 1987), 299–321.

Kargon, R. and Achinstein, P. (eds.), 1987. *Kelvin's Baltimore Lectures and modern theoretical physics: historical and philosophical perspectives*, Cambridge, MA: MIT Press.

Knudsen, O. 1985. 'Mathematics and physical reality in William Thomson's electromagnetic theory', in [Harman 1985], 149–179.

Larmor, J. 1904. 'Lord Kelvin on optical and molecular dynamics', *Nature*, *70* (Supplement), iii–v.

Smith, C. and Wise, M.N. 1989. *Energy and Empire: A biographical study of Lord Kelvin*, Cambridge: Cambridge University Press.

Stokes, G.G. *Papers*. *Mathematical and physical papers*, 5 vols., Cambridge: Cambridge University Press, 1880–1905. [Repr. New York: Johnson, 1966.]

Thompson, S.P. 1910. *The life of William Thomson, Baron Kelvin of Largs*, 2 vols., London: Macmillan.

Thomson, W. *Papers*. *Mathematical and physical papers*, 6 vols., Cambridge: Cambridge University Press, 1882–1911.

Thomson, W. 1872. *Reprint of papers on electrostatics and magnetism*, London: Macmillan.

Wilson, D.B. (ed.) 1990. *The correspondence between Sir George Gabriel Stokes and Sir William Thomson, Baron Kelvin of Largs*, 2 vols., Cambridge: Cambridge University Press.

Wise, M.N. and Smith, C. 1987. 'The practical imperative: Kelvin challenges the Maxwellians', in [Kargon and Achinstein, 1987], 323–348.

CHAPTER 59

HENRI LEBESGUE AND RENÉ BAIRE, THREE BOOKS ON MATHEMATICAL ANALYSIS (1904–1906)

Roger Cooke

These three monographs made the case for the presentation of real-variable analysis developed by a pleiad of brilliant French mathematicians at the turn of the century, including C. Jordan, E. Borel, P. Fatou, M. Fréchet and these two authors. The new theory of measure was especially significant.

Henri Lebesgue, ***Leçons sur l'intégration et la recherche des fonctions primitives,*** first edition (1904)

First publication. Paris: Gauthier–Villars, 1904. viii + 138 pages.

Photoreprint. In *Oeuvres scientifiques*, vol. 2, Geneva: Kundig, 1972, 11–154.

Second edition. Paris: Gauthier–Villars, 1928. xv + 342 pages. [Photorepr. New York: Chelsea, 1950.]

René Baire, ***Leçons sur les fonctions discontinues*** (1905)

First publication. Paris: Gauthier–Villars, 1905. viii + 127 pages.

Photoreprint. Paris: Gauthier–Villars, 1930. Also in *Oeuvres scientifiques*, Paris: Bordas, 1990, 195–327.

Henri Lebesgue, ***Leçons sur les séries trigonométriques professées au Collège de France*** (1906)

First publication. Paris: Gauthier–Villars, 1906. v + 128 pages.

Photoreprint. Paris: Blanchard, 1975.

Related articles: Cauchy (§25, §28), Fourier (§26), Riemann on trigonometric series (§38), Cantor (§46).

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

1 INTEGRALS AND FUNCTIONS IN THE 19TH CENTURY

The concepts of function and integral lie at the heart of much of the development of analysis in the 19th century. The notion of a function became ever more abstract as the century progressed. It began the century as essentially the concept described by Johann Bernoulli at the beginning of the preceding century: 'an expression formed in any manner from variables and constants'. The undefined phrase 'in any manner' was to be understood in a particular context of mathematical objects. It would not have occurred to Bernoulli's contemporaries, for example, to regard the price of a load of grain as a function of its volume or weight. The rules for forming expressions were restricted to algebraic formulas and the family of elementary transcendental functions defined by exponential, logarithmic, and trigonometric functions. The transformation of this notion into its more modern, abstract form is the result of influences from many areas, such as complex analysis, geometry, and algebra.

1.1 Trigonometric series

One of the main influences, whose effects can be seen at many crucial moments, is the representation of quantities by trigonometric series. To explain why trigonometric series forced an enlargement of the concept of a function when power series did not, one need only note that if a power series

$$\sum_{n=0}^{\infty} c_n (z - z_0)^n \tag{1}$$

converges at any point $z = w$, then it converges at all points z such that $|z - z_0| < |w - z_0|$ and represents an analytic function inside that region. Convergent power series of this type allow all the ordinary operations of algebra and calculus to be performed termwise. As long as mathematicians allowed only these operations, they were effectively considering only convergent power series. G.W. Leibniz's discovery of power-series representations for exponential, logarithmic, and trigonometric functions provided grounds for believing that all mathematical problems of any interest could be solved using such functions. Thus, except for a few anomalies noted by Euler and Daniel Bernoulli in the 1740s, where trigonometric representations seemed more natural, Bernoulli's definition seemed adequate at the beginning of the 18th century.

The anomalies just mentioned from the mid 18th century gave an advance glimpse of the problems that would have to be faced if trigonometric series became a major tool for solving differential equations. One way of describing the situation is to say that a trigonometric series could represent different analytic functions in different domains. In studying the equation satisfied by the displacements of a vibrating string Daniel Bernoulli (1700–1782) had considered the problem of finding a function $u(x, t)$ that would satisfy the one-dimensional wave equation $\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}$ together with the appropriate initial and boundary conditions that would represent the case of a string clamped at two points, stretched into an arbitrary shape, and released from a position of rest. These restrictions are expressed (in modern notation) by the boundary conditions $u(0, t) = 0 = u(L, t)$ and the initial conditions $u(x, 0) = f(x)$, $\frac{\partial u}{\partial t} = 0$. Daniel Bernoulli claimed that the coefficients c_n could be

chosen so that the function

$$u(x,t) = \sum_{n=1}^{\infty} c_n \cos(n\pi ct/L) \sin(n\pi x/L) \tag{1}$$

would satisfy all these conditions. (He did not explicitly state the time terms.) The individual terms certainly satisfy all of them except the equation $u(x,0) = f(x)$. Bernoulli thought one could choose the c_n so that $f(x) = \sum_{n=1}^{\infty} c_n \sin(n\pi x/L)$. His former colleague at the Petersburg Academy of Sciences, Leonhard Euler (1707–1783), was not so sure, since the right-hand side always represented an odd function of period $2L$. Suppose the formula by which the function $f(x)$ was defined did not satisfy these two conditions. How could this representation be valid over only *part* of the domain and represent not the function itself but its odd periodic extension outside the domain? To be sure, a power series such as the geometric series may represent its formula only over a limited range, but that is because the series itself diverges outside that range. Such is not the case with the Fourier series. The notion of 'forming an expression from variables and constants' turned out to be less clear than had been believed, since two ways of doing so could coincide over one interval and differ over another.

In the early 19th century Joseph Fourier (1768–1830) made use of trigonometric series representations to discuss the diffusion of heat in physical bodies and also introduced integrals of the type

$$u(x,t) = \int_0^{\infty} f(s) e^{-s^2 t} \cos(xs)\, ds \tag{2}$$

(notation slightly updated) in order to satisfy the heat equation $\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$. What made these trigonometric series and integrals the object of intense scrutiny was their ability to represent *different* functions, functions defined piecewise, as we would now say. Moreover, finding the coefficients of the trigonometric series to represent a function, as Fourier showed, meant integrating the product of the function with a trigonometric function. (Fourier gave an alternate way of computing the coefficients; see §26.5.) How was this to be done? Integration had been regarded as the inverse of differentiation. Integrating a power series was a matter of invoking a simple formula. But in these new applications the function had somehow to be presented. One could not, without circularity, use the series to present the function and then use the function to compute the coefficients of the series. Such a problem did not arise when only power series were used to represent functions, since the coefficients of the series could be determined, usually from a differential equation that the function had to satisfy. Indeed, Karl Weierstrass argued that one could use a differential equation or a system of such equations as the definition of an analytic function. But the method of using trigonometric series to solve equations was based on getting terms that satisfied the differential equation with *arbitrary* coefficients and then using the initial or boundary conditions to find the coefficients. Hence a new concept of integration was needed as well. As long as functions were explicit formulas, integration could simply be the inverse of differentiation; but the kind of initial positions for a vibrating string and initial temperature distributions could obviously not be expected to be one of the familiar functions. A way would need to be found to integrate such functions.

1.2 Functions and their integrals

As A.L. Cauchy (1789–1857) discovered, the need to study functions of a complex variable also demanded a new definition of the integral. He provided such a definition in 1823, treating the integral as a limit of finite sums rather than the sum of infinitesimal products (§25.4). Yet he did not seek the most general conditions under which the limit of the approximating sums could exist, but contented himself with integrating continuous functions, another concept whose definition he reformed. As we now know, when the integral is restricted to this class of functions, no modifications are needed in the fundamental theorem of calculus. Before Cauchy a function had been called continuous if it was defined by a single analytic formula of the type envisaged by Bernoulli. In 1821 Cauchy replaced this definition with the requirement that '$f(x+a)-f(x)$ decreases indefinitely with a', an informal way of stating what is still today the definition of continuity. It should be remembered, however, that the most general functions anyone had considered up to that time were sums of trigonometric series. No one at the time could have guessed how much arbitrariness such a function could exhibit in its behavior. Efforts during the 1820s were concentrated on proving that the Fourier series of a function really does represent the function. The most notable result was due to J.P.G. Lejeune-Dirichlet (1805–1859), who proved in 1829 that convergence holds for a function having only a finite number of maxima and minima and a finite number of discontinuities.

It occurred to Dirichlet that even when the Fourier series of a function converged to the function, there might nevertheless be *other* trigonometric series that also converged to it. Dirichlet suggested that Riemann study this question. While engaged in this study in 1854, Bernhard Riemann (1826–1866) examined Cauchy's notion of integral in more detail, especially the conditions under which the approximating sums would have a limit, so that the integral could be defined. The condition he gave made it possible to integrate many more functions than analysts had ever imagined could be of use. For example, an integrable function could be discontinuous on a dense set of points. Riemann offered the example of the function

$$f(x) = (x) + \frac{(2x)}{4} + \frac{(3x)}{9} + \cdots, \tag{3}$$

where $(a) = a - n$, n being the integer closest to a. (When a is half of an odd integer, (a) is set equal to zero.) When this work was published in 1867, after Riemann's death, it appeared that the integral defined by Riemann was sufficiently general for all purposes of analysis (§38). It was at least possible to prove that if a trigonometric series converges to an integrable function at every point, it must be the Fourier series of that function. That is, its coefficients must be given by the integral formulas used to calculate the coefficients of the Fourier series.

However, this very result pointed to a difficulty with the Riemann integral. One of Fourier's derivations of the formulas for the coefficients was purely formal; it involved assuming that the series could be integrated term by term. That assumption, if true, would automatically have guaranteed the uniqueness of the series. The delicate analysis that Riemann had to perform in order to prove uniqueness underlined the need for conditions under which such termwise integration was possible. Uniform convergence, which is still taught

to undergraduates as a sufficient condition, was of use only in cases where the sum of the series is itself continuous; but the main usefulness of Fourier series occurred precisely in cases when the sum was not continuous.

In demonstrating the possibility of integrating discontinuous functions, Riemann had created the need to classify functions according to their degree of discontinuity. A step in that direction was taken by Gaston Darboux (1842–1917) in an article on discontinuous functions [Darboux, 1875]. He refined the Riemann integral by considering its upper and lower sums (now called the Darboux sums). Relying on the existence of discontinuous integrable functions, Darboux was able to construct continuous functions that do not have a derivative at any point. This paper marks an abrupt departure from geometric intuition to verbal reasoning based on general premises. As Darboux said, 'in the presence of such singular propositions' it was necessary 'to bring the greatest possible rigor to the proofs and allow only the best-established propositions' to be invoked. These nowhere-differentiable continuous functions, and their opposite number—discontinuous functions that nevertheless have the intermediate-value property—formed what Lebesgue was later to refer to as a 'sort of chamber of horrors' [Gispert, 1995, 46]. Unlike 20th-century mathematicians, Darboux had no intention of introducing such monstrosities into analysis; he used them only to show by contrast what the proper objects of analysis were.

By redefining the notion of continuity Cauchy had provided mathematicians with a much larger class of functions than had previously been available. It was, however, not clear precisely what functional relations could exist between variables. Even late in the 19th century, Weierstrass never accepted Cauchy's abstract approach to analytic function theory, insisting that the proper definition of analytic functions was through power series (whose coefficients, as noted above, could be generated by the requirement of satisfying a differential equation). Weierstrass's point of view was that in order to verify that a function is continuous or has a derivative, one must first have a definition of the function. Where is that definition to come from, if not from a power series? As he said in 1885, 'No matter how you twist and turn, you cannot avoid using some analytic expressions'. To "tame" Cauchy's abstract notion of continuity Weierstrass showed in 1885 that any continuous function on a finite interval could be uniformly approximated by a polynomial, and, if periodic, by a trigonometric polynomial of the same period. Since all *physical* quantities are presented with some unavoidable error of measurement, he had effectively shown that there was no loss of generality in thinking of any variable as the sum of a series of polynomials or trigonometric functions. An obvious consequence of his approximation theorems was that the property of being the limit of a sequence of continuous functions was no more general than that of being the limit of a sequence of polynomials or the sum of a trigonometric series. But the question remained: *What kind of* function is the limit of a sequence of continuous functions?

Besides the absence of useful criteria for termwise integration, still other problems arose with the Riemann integral in connection with the fundamental theorem of calculus. In the 1870s Weierstrass and Darboux gave examples of trigonometric series converging uniformly to functions that do not have a derivative at any point. For such functions there could be no question of recovering the function by integrating its derivative. In addition, in 1881 Vito Volterra gave an example of a continuous function having a derivative that was not integrable, thereby raising the question of the proper hypotheses for the fundamen-

tal theorem of calculus to an even more urgent level. Finally, the problem Riemann had set out to solve when he reinvestigated the integral—whether the coefficients of a convergent trigonometric series are determined by its sum when the sum is integrable—remained open in the case when the set of points where the series is not known to converge is not of Cantor's *first kind*.

1.3 Early set theory

In the course of extending Riemann's study of the uniqueness of trigonometric series representations, Georg Cantor (1845–1918) investigated the exceptional sets E such that one can prove uniqueness of a trigonometric series representation without assuming anything about convergence at points of E. Such sets have been called 'sets of uniqueness' (A. Zygmund in the 1930s). Riemann had studied this problem by formally integrating the series twice, thereby producing a continuous function $F(x)$, from which information about the original series was deduced via a generalized second derivative. By looking carefully at this function Cantor showed that such an exceptional set could be any set of *first kind*, that is, one of its derived sets of finite order is empty (§46.2). (The derived set of a set is its set of limit points; by successively repeating the passage to the set of limit points, one gets derived sets of higher order.) The derived set was the first step on the road to a complete classification of sets according to their complexity by counting the number of times one must perform a countable union or intersection in order to reach a given set, starting from the open and closed sets. This classification, although it did not at first seem cogent to many mathematicians, lurked in the background during the process of refining the notions of function and integral. As it finally turned out, the two efforts—to characterize the functions analysis could integrate, and to classify sets, made a very fruitful alliance, each validating the importance of the other.

The geometric interpretation of the integral as an area meant that each generalization of the concept of integral provided an automatic enlargement of the class of plane figures whose areas could be computed and an enlargement of the class of curves whose lengths could be computed. When mathematicians began to think in terms of sets rather than geometric figures, the question whether every linear set could be assigned a length and every planar set an area became a natural one. Thus, the concepts of function and integral were intimately bound up with questions of length and area. The creation of a comprehensive theory of functions that would systematically lay out the permissible definitions of functions and say which ones could be integrated would bring with it a classification of linear and planar sets and their lengths and areas respectively. In order to be of use, such a theory would have to incorporate reasonable criteria for term-by-term integration of a convergent sequence and provide a clear set of hypotheses under which a continuous function is the integral of its derivative. Moreover, since the invention of set theory, the very structure of the real line on which analysis takes place had changed. Where analysts had previously discussed points and intervals, there was now the completely abstract notion of a *set of points*, and there seemed to be no natural limits on the means by which such a set could legitimately be defined.

In this context the Riemann integral was inadequate for the needs of analysis, and analysts began the search for an improved theory. For an integral the desiderata were the

following properties: 1) it should be defined for a sufficiently wide class of discontinuous functions; 2) it should admit an interpretation as 'the area under the curve' (or as volume, in the case of double integrals). An auxiliary problem to be solved was to characterize the continuous functions that have a derivative and define the integral so that such a function is the integral of its derivative.

Some early progress on this last problem was made by Ulisse Dini during the late 1870s, when he showed that a function $f(x)$ such that $f(x) + Ax + B$ is piecewise monotonic for all but a finite number of values of A has a derivative on a dense set. (The reason for including the superfluous constant B is not clear.) Dini had thereby exhibited monotonicity as a possible sufficient condition for a derivative to exist. Hermann Hankel had made an attempt to characterize the discontinuous integrable functions in 1870 by defining a function to be 'pointwise discontinuous' if for each positive number ε the set of points x such that $|f(x+h) - f(x) > \varepsilon|$ for arbitrarily small h contains no intervals [Hankel, 1870]. He believed, and believed he had proved, that such a function met Riemann's sufficient condition for integrability, that is, that the discontinuities of such a function could be enclosed in a set of intervals of arbitrarily short total length.

The problems to be addressed focused on four areas: 1) determining which functions can be the limit of a sequence of continuous functions; 2) determining which functions can be assigned an integral and which sets can be assigned a length, area, or volume; 3) establishing sufficiently nonrestrictive conditions for term-by-term integration of a sequence of functions; and 4) clarifying the conditions under which a continuous function has a derivative of which it is the integral. These were the problems addressed and largely answered in the three works under discussion.

Several mathematicians, including Axel Harnack, Otto Stolz, Giuseppe Peano, and Cantor, introduced the notion of the content of a set in connection with integration [Hawkins, 2001, ch. 3]. The idea was that when a region was partitioned into sufficiently small rectangles, the total area of the rectangles that intersect the boundary of the set should become arbitrarily small. In this way a distinction between measurable and nonmeasurable domains arose. Camille Jordan, in particular, noted the important fact that a finite union of measurable domains should also be measurable.

The one-dimensional version of this work was carried out by Émile Borel in his 1894 dissertation on the convergence of general series of complex functions of the form $\sum_{n=0}^{\infty} A_n(z-a_n)^{m_n}$. The principles used in this work were applied in a monograph [Borel, 1898] that Borel wrote because, as he said in the preface, it was becoming more and more difficult to read research papers knowing only the portions of the theory that were regarded as 'classical'. He specifically mentioned the need for an exposition of the theory of sets, which occupies the first three of the six chapters of the monograph. Chapter 3 of this monograph contains the elementary parts of the theories now called descriptive set theory and measure theory. The descriptive set theory consisted of the definition of closed ('*relativement parfait*') and perfect ('*absolument parfait*') sets, and the proof that every closed set consists of the union of a perfect set and a countable set. The measure theory amounted to the proof that no interval (a, b) can be covered by intervals of total length less than $b - a$. (The proof of this result uses the Heine–Borel theorem.) In this monograph Borel stated axiomatically what he demanded of the concept of measure, specifically that it be finitely subtractive and countably additive.

The only sets that Borel guaranteed measurable at the time, however, were the topologically simplest: the closed sets, and all countable sets, and he was not at all clear in describing just which sets are measurable, except for saying in a footnote that one could easily establish the consistency of his axioms by techniques analogous to those he had used in his earlier arguments. The lack of clarity appeared starkly [1898, 48]. On the one hand, Borel defined sets to be measurable if their measure could be defined by his preceding definition. In modern terms, he was asserting that a set is measurable if it belongs to the smallest class that contains all closed sets and is closed under countable unions and set differences. In his honor, and because of this statement, that class is now called the *Borel sets*. On the other hand, he also said that there might be other sets to which measure could be assigned, and he defined a set to have measure 'at least α' if it contained a measurable set of measure α and 'at most α 'if it was contained in such a set, 'without worrying whether the set is measurable or not'. Thus he also appears to have defined inner and outer measure and what are now called the *Lebesgue-measurable sets*.

To Borel's definition Artur Schoenflies, writing a generally favorable report on the development of set theory for the *Deutsche Mathematiker-Vereinigung* in 1900, objected that 'the question whether a property is extendable from finite to infinite sums cannot be settled by positing it but rather requires investigation' [Hawkins, 2001, 107].

2 THE AUTHORS

Enter now our two authors. Henri Léon Lebesgue, was born on 28 June 1875 in Beauvais. His father worked in a print shop and his mother was a schoolteacher. He studied at the *École Normale* from 1894 to 1897 and remained for further study until 1899. He received the doctorate in 1902 and spent the following year lecturing at the *Collège de France* under a Peccot Foundation Fellowship. In 1910 he became a lecturer at the Sorbonne. During the Great War he worked on ballistic problems. In 1919 he became professor at the Sorbonne and in 1921 professor at the *Collège de France*, where he remained for the rest of his life. He died on 26 July 1941.

René-Louis Baire was born 21 January 1874 in Paris, the son of a tailor. At the age of 12 he won a scholarship that enabled him to study at the *Lycée Lakanal*. In 1892 he entered the *École Normale*, where his written work was quite good. Unfortunately, while proving the continuity of the exponential function during an oral examination, he began to see difficulties with the proof. This lapse damaged his prospects, and he determined to take the examination over. However, it did set him a topic that he found to be of great interest, which guided his research over the next few years. After passing the examination, he won another scholarship for study in Italy, where he met Vito Volterra. For his results on discontinuous functions he received the doctorate in 1899, with a somewhat reserved recommendation by his examining committee, which consisted of Darboux, Appell, and Picard. He taught at Montpellier starting in 1901, and in January/February 1904 he gave the Peccot course at the *Collège de France*, which led to the monograph to be discussed here. When he returned to Montpellier, he underwent a spell of depression, accompanied by esophageal constrictions which made every meal a trial [Dugac, 1976, 309].

In 1905 Baire began teaching at Dijon. The rector of the Dijon Academy noted in 1908 that Baire's health seemed to be fragile, rendering him neuarasthenic and incapable of

teaching with the vigor he would have wished. The esophageal constrictions grew worse over time, and the last 20 years of his life were spent in nearly constant pain. In June 1932 he suffered a particularly bad attack. His sister-in-law came to see him on the shore of Lake Geneva, where he had gone to live; she believed his symptoms were largely nervous ones. On 1 July he was taken to a psychiatric hospital some 100 kilometers away, where he died on 5 July [Dugac, 1976, 313].

3 LEBESGUE'S *LEÇONS SUR L'INTÉGRATION ET LA RECHERCHE DES FONCTIONS PRIMITIVES* (1904)

Lebesgue worked out his theory of measure and integration in a series of research announcements in the *Comptes rendus* between 1898 and 1901. Although his work was published before Baire's monograph, the original research on which both were based was essentially simultaneous, and Lebesgue took a keen interest in Baire's study of the representation of discontinuous functions, on which he also wrote a paper. Lebesgue's approach was very hesitant, as Cantor's set theory had by no means 'caught on'. In particular, Charles Hermite had opposed the publication of one of his research announcements [Hawkins, 2001, 120–122]. Lebesgue's doctoral thesis, entitled *Intégrale, longueur, aire*, was published in the *Annali di matematica* in 1902. During the academic year 1902–1903 he was invited to lecture at the *Collège de France* which gave him the opportunity to reflect further on his previous work and settle a few open questions involving the fundamental theorem of calculus. The fruit of that year was the monograph we are about to discuss. Its contents are summarised in Table 1. The later edition of 1928 expanded to 11 chapters.

In the preface Lebesgue explained the need for the new, abstract definition of a function. As he wrote,

> One may well wonder, it is true, whether there is any interest in studying such complications, and whether it might not be better to limit ourselves to the study of functions that require only simple definitions. Such an approach has only advantages in the case of an elementary course; but, as will be seen in the following lectures, if we wished to limit ourselves always to these good functions, we would have to give up on the solution of a number of easily stated problems that have been open for a long time. It was the solution of these problems, rather than a love of complications, that caused me to introduce in this book a definition of the integral that is more general than that of Riemann and contains the latter as a special case.

After explaining that his new integral was just as simple as that of Riemann and provided simpler proofs of many theorems, even those stated only for the Riemann integral, Lebesgue continued,

> As applications of the definition of the integral I have studied the problem of finding primitive functions and rectifying curves. To these two applications, I would have liked to add another, of great importance: the study of the expansion of functions in trigonometric series. However, while teaching the course I was able to give only such incomplete indications of this topic that I did not consider it worthwhile to reproduce them here.

Table 1. Contents by chapters of Lebesgue's *Leçons sur l'intégration*.The pages of the two editions are indicated. The square brackets for chapter VII enclose the extra part of the title in the second edition.

Ch.	**1st**	**2nd**	**Title**
I	1	1	The integral before Riemann.
II	15	15	Riemann's definition of the integral.
III	36	36	Geometric definition of the integral.
IV	49	49	Functions of bounded variation.
V	64	68	Finding primitive functions.
VI	85	92	The definite integral found using primitive functions.
VII	98	105	[The definite integral of] Summable functions.
Note	131		On sets of numbers. [End 136.]
VIII		141	The indefinite integral of summable functions.
IX		174	Finding primitive functions. The existence of derivatives.
X		202	Totalization.
XI		252	The Stieltjes integral.
Note		314	Appendix on transfinite numbers. [End 340.]

The implied promise in this last paragraph was to be fulfilled two years later with the publication of Lebesgue's monograph on trigonometric series (section 6).

The inclusion of 'primitive functions' in the title of the work emphasizes one of its major aims: to resurrect the fundamental theorem of calculus as a basic tool in the context of an integral more general than those previously considered. In the historical résumé of the first chapter Lebesgue took the original definition of the integral to be the inverse of differentiation. He looked very carefully at the exact usage of the term *function* by his predecessors, saying that

> in reality the correspondences considered by Cauchy remained those that could be defined using analytic expressions. For, after defining functions, Cauchy adds, 'Functions are called *explicit* if the equation connecting x and y is solved for y and *implicit* otherwise'. The fact that the correspondences are defined by analytic expressions is never used in Cauchy's reasoning, so that the properties obtained by Cauchy carry over directly, along with their proofs, to functions satisfying Riemann's definition.

In a footnote at this point he adds:

> I do not mean that Cauchy's definition is less general than Riemann's: at present no Riemannian function is known that does not have an analytic representation. All I am saying is that, if there do exist functions satisfying Riemann's definition but not Cauchy's, they are not excluded from Cauchy's arguments.

Looking back at Lebesgue's language from the perspective of another century, we must be careful to remember that he used the word *analytic* in a special sense that he himself was to define later; in particular, he did *not* mean analytic in the sense of representation by power series.

In Chapter 2 Lebesgue gave a detailed discussion of the Riemann integral, including the conditions for integrability of a function. Riemann had shown that the following condition was necessary and sufficient: *for each* $\varepsilon > 0$ *there exists a partition of the interval of integration such that the total length of the intervals on which the oscillation of the function is larger than* ε *can be made as small as desired*. Lebesgue showed that this condition could be more elegantly stated by saying that the set of discontinuities of the function formed a set of measure zero.

Lebesgue followed this analytic discussion with a geometric discussion in Chapter 3, first defining what is now called the Jordan content of a planar region and proving that a nonnegative function is Riemann integrable if and only if the region below its graph is 'squarable', that is, has a well-defined Jordan content.

Lebesgue's aim, however, was to restore the fundamental theorem of calculus to the place it had lost as a result of the examples of Darboux and Volterra. To that end, Chapter 4 was devoted to the study of functions of bounded variation and the rectification of curves. With that background, Lebesgue took up the search for primitive functions (functions having a given function as derivative) in Chapter 5. He began by noting that the indefinite integral of a Riemann-integrable function is of bounded variation and has the integrand as its derivative at each point where the integrand is continuous. He then called attention to Riemann's example of an integrable function whose discontinuities are dense, showing that the indefinite integral of this function had a derivative equal to the function at all points except rational numbers whose denominators (in lowest terms) are even. From the conclusion of Chapter 2 and the fact that the indefinite integral has a derivative at each point where the integrand is continuous, as Lebesgue pointed out, it followed that the nondifferentiable continuous functions of Weierstrass and Darboux could not be the indefinite integrals of Riemann-integrable functions.

The problem of differentiability, it will be recalled, had been addressed by Dini, who had noticed its connection with monotonicity. It was Dini who introduced the four *derivates* of a function $f(x)$, that is, the upper and lower limits of the difference quotient

$$\frac{f(x+h)-f(x)}{h} \tag{4}$$

as h tends to zero from the right or left. Lebesgue studied the question of whether and how these four derivates determine the function $f(x)$. He showed that if all four numbers were bounded, then the function was completely determined (up to an additive constant) if its upper right derivate was prescribed except on a set of measure zero. Defining the mean value of the function $f(x)$ to be

$$\lim_{h\downarrow 0} \frac{1}{2h} \int_{x_0-h}^{x_0+h} f(x)\,dx \tag{5}$$

at each point x_0 where this limit exists, Lebesgue showed that an integrable function is a derivative (at every point) if and only if the mean value exists at each point and is equal to the function. It was precisely at this point that he considered the Riemann integral inadequate. For, as he said, it was always possible to find a primitive whose derivative is $f(x)$ at each point where this function is continuous (either the upper or lower Riemann integral has this property). But, he said, the problem was indeterminate, since two distinct primitives did not necessarily differ by a constant. To make the problem determinate, he simply required the primitive to have bounded derivates. The solutions already noted still exist, but now, because of his earlier theorem, any two primitives differ by a constant. Having given more than fair coverage to the Riemann integral with these lengthy preliminaries, Lebesgue was at last ready to present his own integral.

In Chapter 6 Lebesgue approached the problem through primitive functions, defining a function to be *summable* if it was the derivative, except on a set of measure zero, of a function having bounded derivates. He showed, with an extra hypothesis, what is now referred to as the monotone convergence theorem: *If a sequence of integrable functions increases to an integrable function, the integral of the limit is the limit of the integrals of the terms of the sequence*.

Only in his seventh and last chapter did Lebesgue finally develop his integral as it is now generally presented, by defining measure and showing that it is countably additive (see, for example, [Burkill, 1951]). The route followed was very similar to many modern presentations, and was generalized later by C. Carathéodory to abstract measure spaces without the need to introduce any essentially new ideas.

4 RECEPTION OF LEBESGUE'S BOOK

The thoroughness with which Lebesgue had investigated the connection between length, area, and primitive functions, and the great generality that his integral allowed in handling termwise integration were powerful factors in favor of his approach to analysis. As already noted, however, many other mathematicians had developed generalized integrals, some (in particular, W.H. Young) very close to the one Lebesgue himself had created. The very generality of his integral was at first considered a disadvantage. Two decades later, when his integral had become an established part of the curriculum and his book required a second edition, Lebesgue recalled the reception his work had met in the early days:

> Although the first edition seemed to some audaciously and gratuitously filled with rather scandalous novelties, it was the work of a timid man who had dedicated six of the seven chapters that he wrote to an exposition of earlier work before embarking on the work that was considered revolutionary. That was done not as the machinations of a propagandist seeking recruits for the revolution, but only to reassure himself. Indeed, he believed, and still believes, that in order to do anything useful one must travel over paths that have been opened by previous work, that doing otherwise carries too much risk of creating a science having no relation to the rest of mathematics.

For this second edition Lebesgue updated what he had written earlier. (At the time of the earlier writing, for example, he had not known whether any non-measurable sets existed;

Giuseppe Vitali constructed an example of one in 1905.) He also added four new chapters, mostly to provide an exposition of what had been learned about absolutely continuous functions and the Lebesgue–Stieltjes integral.

What caused Lebesgue's integral to 'catch on' was its usefulness in connection with much other work, especially Hilbert's work on integral equations, which required square-integrable functions. When combined with the work of Maurice Fréchet, who had created the notion of an abstract metric space, Lebesgue's integral generated a whole class of spaces, the L^p spaces, that became basic objects for harmonic analysis, thereby conferring immortality on Lebesgue's creation. A sketch of that development will be given in connection with Lebesgue's book on trigonometric series, to be discussed below. First, though, we turn to the related work of Baire.

5 BAIRE'S *LEÇONS SUR LES FONCTIONS DISCONTINUES* (1905)

Although the French mathematicians of the 1870s showed little interest in the bizarre counterexamples of Darboux, and Darboux himself soon stopped studying them and never used them, there was interest in such examples, especially in Germany, where the Weierstrassian tradition of rigor was ascendant. It was two decades later when this topic attracted the interest of the young René Baire, who noticed what had already been pointed out by Thomae, K.H.A. Schwarz, and Dini: a function of two variables that is continuous in each variable separately need not be continuous in the two variables jointly. But where Darboux had attempted only the most rudimentary classification of discontinuous functions into those that were integrable and those that were not, Baire took full advantage of the set theory that had been developed in the meantime and used the set of discontinuities as an indicator of other properties of a function. In particular he considered the question of which discontinuous functions can be the limit of a sequence of continuous functions and thereby arrived at the famous Baire classification of functions. In contrast to the reception Darboux had encountered, Baire found a great deal of interest in his work. As Gispert [1995] says, 'In the early years of the 20th century these "most general functions" were no longer the marginal objects they had been in the 1870s. Thirty years after [the work of] Darboux, the theory of sets and the powerful new techniques it provided for studying sets of points had begun to shift the standards in the theory of functions'. An indication of this interest is shown by the fact that Émile Picard (1856–1941) found Baire's thesis worthy of mention in a lecture given on 5 July 1899 at Clark University in Worcester, Massachusetts [Picard, 1905, 20]. Three months earlier Picard had written a report on Baire's thesis [Dugac, 1976, 339–341], and he now said publicly what he had written privately:

> Among the latest work on these delicate questions, I would like to spend a minute discussing a memoir of M. Baire containing some unusual results. The author has succeeded in finding a necessary and sufficient condition for a function $f(x)$ of a real variable to be represented by a simple series of polynomials; the statement of this result requires certain notions on the discontinuity of a function with respect to a set of points: a function may be pointwise or totally discontinuous relative to the set. The condition obtained is that the function be pointwise discontinuous with respect to every perfect set. [See below for

the definition.] M. Baire also poses an unusual question about linear partial differential equations. Consider the equation

$$\frac{\partial f}{\partial x}+\frac{\partial f}{\partial y}=0; \qquad [(6)]$$

if I were to ask which functions satisfy this equation, I would no doubt be told that only functions of $x-y$ satisfy them. M. Baire is not completely sure of that; he notes that the theory of change of variables assumes the continuity of the derivatives used; if one assumes only that the derivatives $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ of the unknown function f exist, one cannot perform the classical change of variables. A delicate analysis is required to establish that a function f, assumed to be continuous with respect to the variables x and y jointly and to satisfy (6), is a function of $x-y$. The conclusion is in doubt if f is continuous only with respect to the two variables individually.

Standing on the brink of the 20th century and seeing the beginnings of the new descriptive set theory and measure theory, Picard rendered the following prophetic judgment [Picard, 1905, 24]:

> Some questions are of purely philosophical interest and will probably never have the least utility for mathematics, for example, to know whether priority belongs to cardinal or ordinal numbers, that is, whether the idea of number proper is anterior to that of rank or vice versa. But, the matter is different in other cases. Thus, it is probable that M. Cantor's theory of sets, which we have already encountered twice, is about to play a useful role in problems that were not posed expressly to be an application of the theory. Therefore let us not begrudge the arduous work on the idea of number and function, for the theory of functions of a real variable is the true basis of mathematical analysis.

The useful role that Picard saw for set theory was already beginning to take shape as he wrote these words. His prediction may have been evoked by noticing the heavy use that Baire was making of Cantor's set theory. Baire's book is summarised in Table 2.

Mathematical writing from the early 20th century frequently seems to treat discontinuity as if it were a positive property, rather than merely the absence of continuity. In the preface of his book, for example, Baire asked, 'Is it not the duty of the mathematician to begin by

Table 2. Contents by chapters of Baire's book.

Chapter	**Page**	**Title**
I	1	The elementary study of discontinuous functions.
II	23	Well-ordered sets and transfinite numbers.
III	50	Subsets of the line.
IV	69	Functions of one variable.
V	99	Functions of n variables. [End 127.]

studying *in abstracto* the relations between these two concepts of continuity and discontinuity, which, while mutually opposite, are intimately connected?'. A few years later the Moscow mathematicians N.V. Bugaev, P.A. Florenskii, D.M. Egorov, and N.N. Luzin were writing similarly about developing a theory of discontinuous functions *in parallel with* the theory of continuous functions, as if the two things were opposite in the sense that the two gloves of a pair are opposite.

The five chapters of Baire's little masterpiece are very single-mindedly aimed at giving a complete answer to one simple question: *What is a necessary and sufficient condition for a function of a finite number of real variables to be the pointwise limit of a sequence of continuous functions?* Until the last chapter, Baire restricted himself to finite-valued functions of one variable. Only after the basic ideas were made clear in this 'tamer' context did he relax the restrictions and consider extended-real-valued functions of several variables. Much of the material in this book is now classical; some of it indeed constitutes an exposition of the work of Cantor. Starting with simple examples of discontinuous functions that are the sums of series of trigonometric functions in Chapter 1, Baire introduced the basic concept of set theory with which Cantor had begun: the derived set. Exactly as Cantor had done with sets of uniqueness, Baire showed that a function whose set of discontinuities is of Cantor's first kind is the limit of a sequence of continuous functions. He then devoted Chapter 2 to the introduction of transfinite ordinal numbers in order to define the derived sets of infinite order. Having developed the necessary ordinal arithmetic, he returned to the definition of derived sets in Chapter 3, including here the examples of nowhere-dense perfect sets that are now a classical part of the real analysis curriculum and concluding this chapter by proving that every closed subset of the real numbers is the disjoint union of a perfect set and a countable set.

With this now classical material established, Baire returned finally to the main question in Chapter 4, proving that the limit of a sequence of continuous functions of one variable is continuous at all the points of a dense set. Such a function is said to be *pointwise discontinuous*, yielding the rather strange-sounding implication that a continuous function is pointwise discontinuous. This notion had been introduced in [Hankel, 1870], who had defined the jump of a function at a point in a manner essentially equivalent to what is now called the oscillation of the function at the point. He defined a function to be pointwise discontinuous if the set of points where its jump is larger than each fixed positive number is nowhere dense. The opposite of pointwise discontinuity for Hankel was total discontinuity—discontinuity at a dense set of points, as in the example given by Riemann. It follows from his definition that the set of discontinuities of a pointwise-discontinuous function forms what is now called a set of first Baire category (introduced by Baire in Chapter 4), and hence that its complement is dense. Like Darboux a few years later, Hankel believed that the proper business of analysis was confined to the comparatively well-behaved pointwise-discontinuous functions. Baire's work provided some support for this position.

In Chapter 4 Baire showed that this condition must also hold on every perfect subset of the line. He then proved the sufficiency of pointwise discontinuity, but only for functions assuming only the values 0 and 1 (in other words, the characteristic functions of sets). In the final chapter, as mentioned above, the sufficiency and necessity of the condition is shown for functions of several variables, whether finite- or infinite-valued.

In this work Baire had shown by implication what kinds of functions could be represented as the sum of a series of continuous functions. Since representations by orthogonal polynomials and trigonometric polynomials were becoming more numerous every day, he had performed an important service by exhibiting a structural characterization of the class of functions that one would need to integrate in order to obtain the coefficients of these representations. The simple example of a Cantor set of positive measure, which is pointwise discontinuous on every perfect set, yet not Riemann integrable, shows that the Riemann integral is not adequate even for that purpose, even waiving the requirements of the fundamental theorem of calculus and simple conditions for termwise integrability of a sequence.

6 RECEPTION OF BAIRE'S BOOK

Perhaps because he thought no one would be interested in functions too general to be represented as sums of trigonometric series, Baire had omitted from his monograph one of the chapters of his thesis, namely the classification of all functions into a hierarchy indexed by the countable ordinal numbers, each class consisting of the functions that are limits of sequences of functions of lower index, but not themselves equal to any such function. As Picard noted in the lecture quoted above, Baire showed definitively that a function (having possibly infinite values at some points) is the pointwise limit of a sequence of continuous functions if and only if it is pointwise-discontinuous on each perfect set. This property came to be known as the *Baire property*, and formed an important focus of research for the founder of the Russian school of descriptive set theorists, Luzin.

Lebesgue, Baire's contemporary and to some degree rival, took a deep interest in the questions that Baire had studied. He published a long memoir in order to clarify the whole subject [Lebesgue, 1905]. Lebesgue defined the *analytically representable functions* to be those that could be constructed by a countable number of repetitions of algebraic and pointwise limit operations starting from the polynomials (or continuous functions). While admitting that only the simplest of these had actually been used in the mathematical literature, he showed that there were non-analytic functions even among those that are Riemann-integrable. Thus it appeared that integration could lead to functions of seemingly unimaginable complexity. Alternatively, one could look at the question from the other side and assert, as some mathematicians have done, that the integral was 'overdesigned' for any applications it might have. Descriptive set theory was 'caught in the middle'. All the functions analysts were using belonged to the first Baire class, and the ingenuity of Luzin's students was taxed to produce explicit examples of functions belonging to the third or fourth Baire class. On the other hand, the example of the characteristic function of a non-Borel-measurable subset of Cantor's set, which is continuous on the complement of the Cantor set, and hence Riemann integrable, shows that even Riemann integration is capable of dealing with functions that do not admit an analytic representation. (However, it is easily shown that every Lebesgue-measurable function is equal, except on a set of measure zero, to a function that belongs to the second Baire class.)

The importance of Cantor's theory of sets in all this work is obvious, but one also can look at the whole situation from the opposite side. An important early objection to Cantor's

work was that it was all intricate philosophical machinery, from which nothing of mathematical importance could be deduced. Hermite, a model of politeness and tact, expressed essentially this thought to Gustav Mittag-Leffler, when the latter asked him to supervise the translation of Cantor's work into French in 1883. And, as Picard's 1899 lecture indicated, the theory had by no means proved its usefulness to the French mathematicians by the end of the 19th century. The work of Borel, Lebesgue, and P. Fatou did much to provide the important applications that the theory needed in order to gain respectability.

7 LEBESGUE'S *LEÇONS SUR LES SÉRIES TRIGONOMÉTRIQUES* (1906)

A new theory, even an elegant one, will attract only temporary interest unless it solves some interesting problems. Unresolved questions regarding trigonometric series representations had played a large role in driving real analysis to higher levels of abstraction; and these questions were, as Lebesgue and others discovered, a rich field for applications of the Lebesgue integral. Although Lebesgue had been working on such problems while giving his first Peccot course, he had not found time or space enough to give a thorough exposition along with his lectures on the integral. Lebesgue began the process of applying this integral to trigonometric series when he gave a second Peccot course in 1904–1905; the notes from this course provided the material for his second book, a 125-page gem full of subtle and illuminating facts about trigonometric series representations. The book, divided into an introductory section and five chapters, is summarised in Table 3.

The introduction summarizes the material on the Lebesgue integral that will be needed for the applications, in particular the dominated convergence theorem for a finite interval and the fact that translation is continuous, as we would now say, in the metric of Lebesgue-integrable functions.

Much of Lebesgue's first chapter is historical, reviewing the use of trigonometric series by Euler in astronomy, by Daniel Bernoulli in mechanics, and by Fourier in heat diffusion. As the title indicates, this chapter is primarily concerned with methods of finding the coefficients of a trigonometric series to represent a given function. Lebesgue reviewed carefully the Euler–Fourier method of computing the coefficients by integration, an 18th-century method of interpolation used by A. Clairaut and J.L. Lagrange, and Fourier's power-series

Table 3. Contents by chapters of Lebesgue's *Leçons sur les séries trigonométriques*.

Ch.	**Page**	**Title**
	1	Introduction. Properties of functions.
1	17	Determination of the coefficients of the trigonometric series representing a given function.
2	33	Elementary theory of Fourier series.
3	55	Convergent Fourier series.
4	84	Arbitrary Fourier series.
5	110	Arbitrary trigonometric series. [End 125.]

method, which involved sets of linear equations in infinitely many variables. He then defined Fourier series, for summable (Lebesgue-integrable) functions only, as series whose coefficients are given by the Euler/Fourier formulas, using the tilde sign, which he attributed to Hurwitz, to write

$$f(x) \sim \frac{1}{2}a_0 + (a_1 \cos x + b_1 \sin x) + (a_2 \cos 2x + b_2 \sin 2x) + \cdots, \tag{7}$$

which he said could be read as '$f(x)$ has the Fourier series $\frac{1}{2}a_0 + (a_1 \cos x + b_1 \sin x) + \cdots$'. For functions having nonabsolutely convergent integrals, which had been considered by earlier authors, he referred to the corresponding series as *generalized Fourier series*. The crucial point, he said, was to determine whether the tilde sign in (7) could be replaced by an equals sign. Despite physical arguments intended to justify this replacement, Lebesgue said, no such argument could replace a mathematical proof.

Chapter 2 takes up the problem of convergence of Fourier series. Starting with a number of examples of particular series, Lebesgue arrives at what appears to be a rather weak result, namely that a function having a derivative of bounded variation on the intervals of some partition of the interval of periodicity has a convergent Fourier series. Weak as it appears, this result allows Lebesgue to deduce the Weierstrass approximation theorem mentioned above, and to prove that for any continuous function on the unit circle in the complex plane there exists a harmonic function in the unit disk having the given function as its boundary values (the Dirichlet problem). He concludes this chapter with a proof that the Fourier series of a function having right- and left-hand limits at a point cannot converge to any value other than the average of those limits and shows finally that the solution of the Dirichlet problem is unique.

Convergence continued to be the main topic in Chapter 3. Lebesgue proves the now-standard result that the Fourier coefficients of an integrable function tend to zero. He makes frequent use of the expression

$$\int_{\delta}^{\frac{\pi}{2}} \left|\psi(t+\delta) - \psi(t)\right| dt, \tag{8}$$

which plays a role in the most delicate of the standard criteria for convergence of a Fourier series, known as *Lebesgue's condition*. Lebesgue presents a variety of such criteria, for which he cited criteria of Dini, Jordan, Dirichlet, and R. Lipschitz as precedents. He then gives the example of a summable but not Riemann-integrable function representable by a Fourier series. In a section of 'miscellaneous applications' Lebesgue gives the inversion formula for the Fourier integral of an integrable function of bounded variation, as well as the Poisson summation formula connecting the Fourier integral of an integrable function on the line with the Fourier series of the periodic function that is obtained by summing the integrable function at evenly-spaced points.

In Chapter 4 Lebesgue considered divergent Fourier series, giving Paul Du Bois-Reymond's example of a continuous function whose Fourier series diverges at certain points and also an example of a Fourier series that converges non-uniformly to a continuous function. He then took up the study of summation methods for divergent series, discussing in detail the methods of S.D. Poisson, Riemann, and L. Fejér. All this material

is now regarded as standard and basic in the study of Fourier series. The next section considers the allowable termwise operations on Fourier series; in particular, Lebesgue proves what is now called 'Parseval's equality' for the Fourier coefficients of a continuous periodic function and speculated whether one could deduce anything about the Fourier series of a product of two functions having convergent Fourier series. He also shows that a Fourier series can always be integrated termwise, whether or not the series converges. The integrated series necessarily converges to the indefinite integral of the function having the given Fourier series. Then follows a section of geometric applications, containing in particular the elegant proof that any simple closed rectifiable curve of length L can enclose an area at most, $L^2/(4\pi)$, equality holding only in the case of a circle (the isoperimetric inequality).

The fifth and final chapter takes up the subject of general trigonometric series, with coefficients given arbitrarily and not generated by any integrable function. Cantor had shown that the coefficients of any trigonometric series that converges on an interval must tend to zero. Even though Kronecker had shown how to deduce all of Cantor's results without the need to prove this result (he showed that the theorems proved without the assumption can be deduced from the same theorems proved with the assumption), Lebesgue could not resist showing that this theorem, now known as the Cantor–Lebesgue theorem, holds even when one assumes only convergence on a set of positive measure. Although this result plays no important role in the uniqueness theory for trigonometric series in one variable, it turned out many decades later that the two-variable analog of the theorem is of crucial importance in the theory of uniqueness for trigonometric series in two variables [Cooke, 1971]. Lebesgue then gave a discussion paralleling, with commentary, that given by Riemann in 1854, and ending with the uniqueness theorem proved by Cantor, and an example (due to Fatou) of a convergent trigonometric series that is not a Fourier series, namely

$$\sum_{n=2}^{\infty} \frac{\sin nx}{\log n}. \tag{9}$$

In the final section Lebesgue discussed a technical point concerning an auxiliary function introduced by Riemann in the course of his proof of uniqueness, suggesting that it would be of interest to relax the restrictions imposed on this function.

8 RECEPTION OF LEBESGUE'S SECOND BOOK

Simultaneously with this work of Lebesgue, Fatou was studying the Dirichlet problem with summable functions rather than continuous functions as boundary values. To Fatou belongs the credit for realizing the importance of square-summability. Given a Fourier series (7) generated by a square-summable function $f(t)$, Fatou considered the function $u(r,t)$ in polar coordinates in the unit disk defined by the convergent series

$$u(r,t) = \frac{1}{2}a_0 + \sum_{n=1}^{\infty}(a_n \cos nt + b_n \sin nt)r^n. \tag{10}$$

Fatou showed that $u(r,t)$ tends to the function $f(t)$ as re^{it} approaches e^{it} along any non-tangential path. It was Frigyes Riesz (1880–1956) who saw the important connection between square-summable functions and the space Hilbert had used in the theory of integral equations. He realized that the uniform metric Fréchet had introduced for the space of continuous functions could be generalized to what is now called the L^2 metric, and he showed that the trigonometric functions formed a complete orthogonal system in this space (the Riesz–Fischer theorem). By 1910 Riesz had introduced the full set of spaces now known as the L^p spaces. A little later G.H. Hardy combined the work of Riesz and Fatou, introducing the subspace of L^p consisting of those functions that are the boundary-values of analytic functions in the unit disk ([Hardy, 1915]: actually, he showed only that the norm of the function is an increasing function of the radius). These spaces are now called H^p spaces. The spaces L^p and H^p proved to be of enormous mathematical interest, and questions involving the convergence of the Fourier series of functions belonging to one or another of them were the subject of hundreds of papers during the 20th century. Perhaps the problem of greatest interest was formulated by Luzin in his 1914 doctoral thesis 'Integration and the trigonometric series'. Luzin conjectured that the Fourier series of a square-integrable function converges except on a set of measure zero. The proof of this conjecture was finally given by L. Carleson in 1965.

9 JOINT EFFECT OF THESE THREE WORKS

The work of Lebesgue and Baire correspond roughly to two aspects of function theory: measure theory, and the descriptive theory. The descriptive theory is in a definite sense the more basic of the two, since measure and integration theory needs to delineate the classes of measurable sets and functions before measure and integration can be defined. For this purpose the Baire classification of functions is useful, as is the hierarchy of Borel sets. However, the full power of these classifications is not needed, since, by an elementary result of measure theory, every Lebesgue-measurable set differs from a set in the second Borel class by a set of Lebesgue measure zero, and every measurable function is equal almost everywhere to a function of the second Baire class. This amount of descriptive theory is included in every course of measure theory, and the further details of Baire's work are often omitted. Of the two areas, the metric (measure) theory has been by far the more influential. Although many areas of mathematical analysis make use of the Lebesgue integral and no analyst can afford to be ignorant of it, the finer points of descriptive set theory are considered a subject for specialists. Descriptive set theorists, however, have made an attempt in recent years to close the gap between the two, showing that their techniques can be used to produce metamathematical theorems ruling out any simple characterization of the sets of uniqueness for trigonometric series [Cooke, 1993; Kechris and Louveau, 1987].

BIBLIOGRAPHY

Borel, É. 1898. *Leçons sur la théorie des fonctions*, Paris: Gauthier–Villars.

Burkill, J.C. 1951. *The Lebesgue integral*, Cambridge: Cambridge University Press (Tracts in Mathematics and Mathematical Physics, No. 40).

Cooke, R. 1971. 'A Cantor–Lebesgue theorem in two dimensions', *Proceedings of the American Mathematical Society, 30*, 547–550.

Cooke, R. 1993. 'Uniqueness of trigonometric series and descriptive set theory, 1870–1985', *Archive for history of exact sciences, 45*, 281–334.

Darboux, G. 1875. 'Mémoire sur les fonctions discontinues', *Annales scientifiques de l' École Normale Supérieure*, (*2*) *4*, 57–112.

Dugac, P. 1976. 'Notes et documents sur la vie et l'œuvre de René Baire', *Archive for history of exact sciences, 15*, 297–383.

Gispert, H. 1995. 'La théorie des ensembles en France avant la crise de 1905: Baire, Borel, Lebesgue … et tous les autres', *Revue d'histoire des mathématiques, 1*, 39–81.

Hankel, H. 1870. *Untersuchungen über die unendlich oft oscillierenden und unstetigen Functionen*, Tübingen (*Dissertation*). [Repr. in *Mathematische Annalen, 20* (1882), 63–112; also in *Ostwalds Klassiker der exakten Wissenschaften*, no. 153 (ed. P.E.B. Jourdain), Leipzig: Engelsmann, 1905, 44–102.]

Hardy, G.H. 1915. 'The mean value of the modulus of an analytic function', *Proceedings of the London Mathematical Society, 14*, 269–277. [Repr. in *Collected papers*, vol. 3, 549–557.]

Hawkins, T.W. 2001. *Lebesgue's theory of integration: Its origins and development*, 2nd ed., Providence: AMS Chelsea Publishing.

Kechris, A.S. and Louveau, A. 1987. *Descriptive set theory and the structure of sets of uniqueness*, Cambridge: Cambridge University Press (London Mathematical Society Lecture Notes, No. 128).

Lebesgue, H. 1905. 'Sur les fonctions représentables analytiquement', *Journal des mathématiques pures et appliquées, (6) 1*, 139–216. [Repr. in *Oeuvres scientifiques*, vol. 3, 103–180.]

Picard, É. 1905. *Sur le développement d l'analyse et ses rapports avec diverses sciences*, Paris: Gauthier–Villars.

Pier, J.-P. 1990. 'Aspects de l'évolution des mathématiques entre 1900 et 1950', in *Travaux mathématiques*, II, Université Luxembourg, 45–161.

Zoretti, L. 1912. 'Les ensembles de points', in *Encyclopédie des sciences mathématiques pures et appliqués*, Tome II, volume 1, fascicule 2, Paris: Gauthier–Villars, 8–241.

CHAPTER 60

H.A. LORENTZ, LECTURES ON ELECTRON THEORY, FIRST EDITION (1909)

A.J. Kox

In these lectures Lorentz presents the principles and some major applications of his electron theory, especially his atomistic theory of electromagnetism that is based on the existence of elementary charged particles ('electrons') that interact with each other and with electromagnetic fields.

First edition. *The theory of electrons and its applications to the phenomena of light and radiant heat. A course of lectures delivered in Columbia University, New York, in March and April 1906*, Leipzig: Teubner, 1909 (*Sammlung von Lehrbüchern auf dem Gebiete der mathematischen Wissenschaften mit Einschluss ihrer Anwendungen*, vol. 29). 332 pages.

Second edition. 1916. 343 pages. [Photorepr. New York: Dover, 1952. Also as *Selected works*, vol. 5 (ed. N.J. Nersessian and H.F. Cohen), Nieuwerkerk a/d IJssel: Palm Publications, 1987.]

Russian translation of the 2nd edition. *Teoriia elektronov i ee primenenie k iavleniiam sveta i teplovogo izlucheniia*, Leningrad and Moscow: GTTI, 1934.

Related articles: Maxwell (§44), Kelvin (§58), Einstein (§63).

1 BIOGRAPHY

Hendrik Antoon Lorentz was born in Arnhem, the Netherlands in 1853. After completing his studies at the University of Leyden he obtained his doctorate in physics in 1875 on a dissertation dealing with the reflection and refraction of light. Two years later he was appointed Professor of Mathematical Physics at the University of Leyden, a position that he would hold for the rest of his career. In 1902 he shared the Nobel Prize for physics with his Amsterdam colleague Pieter Zeeman for their work on the relation between magnetism and light, one of the major successes of which was the discovery (by Zeeman) and the

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

explanation (by Lorentz) of the Zeeman effect, the splitting of spectral lines in several components under the influence of an external magnetic field.

As his career developed, Lorentz became one of the undisputed world leaders in theoretical physics. Typical for his status is that he chaired all of the famous Solvay Congresses that took place during his lifetime. His chairmanship was applauded by all, and led Albert Einstein to call him a 'living work of art'. For a review of his life and work, see [McCormmach, 1973].

2 LORENTZ'S CONTRIBUTIONS TO ELECTROMAGNETISM

If one looks at the development of Lorentz's work, one is impressed in the first place by the wide range of topics on which he worked and to which he made important contributions: for example, kinetic theory, thermodynamics, radiation theory and quantum theory. But among these there is one topic that stands out and that appears again and again in his writings: the theory of electromagnetism. Already in his dissertation one can discern the beginnings of a systematic research program that would unfold in the following years. One of the innovative aspects of the dissertation is that Lorentz applies the relatively new theory of J.C. Maxwell (§44) to the phenomena of reflection and refraction of light. In this view, light consists of electromagnetic transversal waves that propagate in a medium, called the electromagnetic ether. Earlier, light had been treated as waves in a medium, the 'light ether', that had properties similar to those of an elastic solid. Apart from unifying electrical and magnetic phenomena, Maxwell's theory had the advantage that the light ether and the electromagnetic ether became the same substance, doing away with the awkward problem of longitudinal waves, predicted in the elastic-solid theory but never observed.

A second important aspect of Lorentz's approach to electromagnetism is that he treated matter as atomistic: matter consists of small particles, some of which may be charged, surrounded by ether. In so far as they are charged, the particles exert an influence on the ether. This disturbance propagates through the ether and manifests itself as electromagnetic action as soon as it reaches charged or magnetized matter. Thus, a separation between ether and matter is achieved, where matter is the source or the recipient of electromagnetic action and ether serves as the medium in which the action propagates. This separation, only implicit in Lorentz's early work, would become sharper as the theory developed. Whereas at the beginning the ether was still treated as a mechanical system—in a major paper [Lorentz, 1892] derived Maxwell's equations from a Lagrangian principle for the ether—it gradually loses all mechanical properties except one: its immobility. Maxwell's equations are then simply postulated, in a Hertzian vein.

A third characteristic of Lorentz's approach is his introduction of what he himself called the 'hypothesis of a single mobile particle'. First introduced in a treatise on dispersion phenomena of 1879, this hypothesis postulates that atoms can contain many charged particles, at least one of which is harmonically bound to a center. The periods of the vibrations such a particle can perform—and thus the frequency of the radiation emitted by it—are influenced by electromagnetic fields. Thus first the phenomenon of dispersion and later the experimentally discovered Zeeman effect could be explained in a satisfactory way. At first, the name 'ions' was used, but after the identification of the intra-atomic particles with

the charged particles found in cathode rays, the name ‘electron’ became common, and Lorentz’s electromagnetic theory became known as the ‘theory of electrons’.

Of all the assumed mechanical properties of the ether, its immobility was by far the most important and at the same time the most problematic. From astronomy there was strong evidence that the ether was not dragged along by the Earth; in particular, the phenomenon of stellar aberration could only be explained if one took the ether to be stationary. Attempts to find other ether models that could account for the astronomical observations failed. On the other hand, a whole series of electromagnetic experiments, the best known of which is the famous Michelson–Morley experiment, failed to show any indication of the Earth’s supposed motion through the ether. In the latter experiment, an interference method was used to measure the ‘ether wind’, the phenomenon that the speed of light would have different values parallel and perpendicular to the direction of the Earth’s motion. In general, all electromagnetic phenomena should depend on the speed of the Earth through the ether because the Maxwell equations change form when transformed from an ether-based reference frame to one that moves with respect to the ether. In modern words, one says that the Maxwell’s equations are not invariant under Galileo transformations, the transformations that connect frames that move uniformly with respect to each other.

For Lorentz the whole problem of the motion of the Earth through the ether was of utmost importance. Indeed, in an unpublished manuscript from 1900 in which he describes his scientific development, he explicitly mentions that his idea to make a sharp distinction between the ether and matter was inspired by the hope than in this way a consistent theory of optical phenomena in moving bodies could be developed.

In fact it would be the struggle with this particular problem that led Lorentz to develop his most powerful tools in the theory of electromagnetism. In a series of papers he tried to find a general and systematic method to reduce phenomena in moving systems to equivalent ones in resting systems. This led him to his so-called ‘theorem of corresponding states’, which allowed him to explain the negative outcome of several electromagnetic experiments. But the Michelson–Morley experiment defied his efforts. At his wit’s end, he saw no other possibility than the introduction of the so-called contraction hypothesis: all material bodies that move through the ether will be shortened in their direction of motion by a fraction that depends on the speed with which they move with respect to the resting the ether. Lorentz made the hypothesis plausible by postulating a similarity in behavior between electric forces and the forces that keep material bodies together. Finally, in [Lorentz, 1904], he succeeded in drawing up a unified theory in which the contraction hypothesis follows in a natural way from the theorem of corresponding states. It is in this paper that the transformations are derived that connect phenomena in moving systems to corresponding ones in stationary system, the transformations now known as the ‘Lorentz transformation’. The main conclusion of the paper was that it is impossible to detect the motion of the Earth through the ether using electromagnetic (including optical) experiments. Note the timing of Lorentz’s paper: this was one year before Albert Einstein published his special theory of relativity in which the Lorentz contraction and the Lorentz transformation were given a totally different interpretation.

3 THE LECTURES

Having reached this apparent endpoint, and having been invited to lecture at Columbia University in New York in 1906, it is not surprising that Lorentz chose the theory of electrons as his topic and decided to present a comprehensive and didactically composed overview of this, his life's work. He also decided to rework the lectures into a book. He accomplished this task so beautifully that even today it is a joy to study the book. As the lectures are presented, *The theory of electrons* is a masterly exposition of everything the theory could accomplish. The contents are outlined in Table 1.

Lorentz summarised the approach taken in the lectures thus (p. 8):

> If we want to understand the way in which electric and magnetic properties depend on the temperature, the density, the chemical constitution or the crystalline state of substances, we cannot be satisfied with simply introducing for each substance these coefficients, whose values are to be determined by experiment; we shall be obliged to have recourse to some hypothesis about the mechanism that is at the bottom of the phenomena.
> It is by this necessity, that one has been led to the conception of *electrons*, i.e. of extremely small particles, charged with electricity, which are present in immense numbers in all ponderable bodies, and by whose distribution and motions we endeavor to explain all electric and optical phenomena that are not confined to the free ether.

This is the first true exposition of electrodynamics from an atomistic point of view: every discussion starts with the microscopic Maxwell equations, that is, the equations that hold for systems of atoms and electrons. Why is that important? It is for the simple fact that for those systems there are only *two* fields, instead of the usual four ones. Microscopic charged particles are only subject to the electric force **d** and the magnetic force **h** (in Lorentz's notation). Obviously, these fields are strongly fluctuating on the scale of atoms and electrons; to derive the equations for macroscopic bodies an averaging or smoothing operation has to be carried out. It is at that stage that the other well-known fields appear: the electric displacement and the magnetic induction. This microscopic approach and the conceptual

Table 1. Contents by chapters of Lorentz's book.

Chapter	**Page**	**Topics**
1	1	General principles. Theory of free electrons.
2	68	Emission and absorption of heat.
3	98	Theory of the Zeeman effect.
4	132	Propagation of light in a body composed of molecules. Theory of the inverse Zeeman effect.
5	168	Optical phenomena in moving bodies.
	231	Notes [End 329], index.

simplification that it brought about was characterized by Einstein as an ‘act of liberation’ [Einstein, 1957].

After having discussed the fundamentals of his theory, Lorentz guides the reader through its most important applications, such as the absorption and emission of light, the theory of the Zeeman effect, and the propagation of light in molecular bodies. The discussion is crystal clear, with technical details that would interrupt the flow of the argument relegated to a series of notes at the end of the book. The final chapter is devoted to the problem of optical phenomena in moving bodies, and it is here that we are confronted with the conceptual clash between Lorentz’s ether-based approach and Einstein’s special relativity. Lorentz essentially repeats the argument from his 1904 paper, showing that all attempts at an experimental determination of the motion of the Earth through the ether are doomed to fail. He also comments on the relation between his theory and special relativity thus (pp. 229–230):

> His [Einstein’s] results concerning electromagnetic and optical phenomena [. . .] agree in the main with those we have obtained in the preceding pages, the chief difference being that Einstein simply postulates what we have deduced with some difficulty and not altogether satisfactorily, from the fundamental equations of the electromagnetic field. [. . .] Yet, I think something may also be claimed in favour of the form in which I have presented the theory. I cannot but regard the ether, which can be the seat of the electromagnetic field with its energy and its vibrations, as endowed with a certain degree of substantiality, however different it may be from all ordinary matter.

Indeed, the formalism developed by Lorentz is practically identical to Einstein’s results; in particular, the experimental predictions that could be tested at the time were identical. This held in particular for the outcome of measurements of the velocity dependence of the mass of the electron. Ironically, because they made the same predictions, Einstein’s and Lorentz’s theories were commonly grouped together as the ‘Einstein–Lorentz theory’, to distinguish them from other theories of the structure of the electron, such as the one by Abraham. That this name is a misnomer becomes immediately obvious when we look at the foundations of both theories. The theory of electrons is a classical ether-based theory, in which Maxwell’s equations in their usual form are only valid in a reference frame that is at rest with respect to the ether. On the other hand, special relativity postulates the universality of the principle of relativity: the physical equivalence of all reference frames that move uniformly with respect to each other. That it is impossible to make an experimental distinction between two such frames was the *outcome* of the theory of electrons, whereas in special relativity it was *implied* by the principle of relativity.

A final word is needed about Lorentz and the ether. The state of mind that he expressed so clearly in the quotation above would remain with him for the rest of his life. Although he realized that the existence of the ether could not be shown using electromagnetic experiments and that it in fact served no physical purpose whatsoever, he maintained that he could not see how waves propagate in vacuum and how the same vacuum could be the seat of electromagnetic field energy. He felt that one needed a substrate in the same way one needs a peg if one wants to hang up one’s hat. By the time of his death in 1928, Lorentz was one of the few physicists left who held that view.

4 A NOTE ON HIS IMPACT

If a single work epitomizes Lorentz's life's work, it is his *Theory of electrons*. (The second edition of 1916 contained a small number of changes in the main text and some additions to the footnotes and in the notes section.) This book is the endpoint of a life's work, completing what was started as early as 1875 in his dissertation. It is an ironic twist of history that the book appeared just when Einstein's special theory of relativity, which in effect superseded Lorentz's work, was gaining more and more support in the physics community (compare §63). Still, that does not detract from its great value and influence. The approach taken and the methods used in giving microscopic descriptions of electromagnetic phenomena are still valid and have proven to be extremely useful. Modern-day electromagnetic theory, as it is taught at universities, owes much to Lorentz's work. If we are not always aware of this, then it is perhaps because modern physicists have, in Einstein's words, 'absorbed Lorentz's fundamental ideas so completely that they are hardly able to realise to the full the boldness of these ideas and the simplification which they brought into the foundations of the science of physics' [Einstein, 1957].

BIBLIOGRAPHY

Darrigol, O. 2000. *Electrodynamics from Ampère to Einstein*, Oxford: Oxford University Press.

Einstein, A. 1957. 'H.A. Lorentz, his creative genius and his personality', in G.L. de Haas-Lorentz (ed.), *H.A. Lorentz, impressions of his life and work*, Amsterdam: North-Holland, 5–9.

Kox, A.J. 1988. 'Hendrik Antoon Lorentz, the ether, and the general theory of relativity', *Archive for history of exact sciences*, *38*, 67–78.

Lorentz, H.A. 1892. 'La théorie électromagnétique de Maxwell et son application aux corps mouvants', *Archives néerlandaises*, *25*, 363–552. [Repr. in *Collected papers*, vol. 2 (1936), 164–343.]

Lorentz, H.A. 1904. 'Electromagnetic phenomena in a system moving with any velocity smaller than that of light', *Proceedings Koninklijke Akademie van Wetenschappen (Amsterdam)*, *6*, 809–831. [Repr. in *Collected papers*, vol. 5 (1937), 172–197.]

McCormmach, R. 1973. 'Lorentz, Hendrik Antoon', in *Dictionary of scientific biography*, vol. 8, 487–500.

CHAPTER 61

A.N. WHITEHEAD AND BERTRAND RUSSELL, *PRINCIPIA MATHEMATICA*, FIRST EDITION (1910–1913)

I. Grattan-Guinness

In this mammoth work the authors gave a detailed account of mathematical logic and set theory, and argued that 'all', or at least much, mathematics could be built upon it. While the reactions were negative as well as positive, the book helped to stimulate much work on logic and the foundations of mathematics.

First publication. 3 volumes, Cambridge: Cambridge University Press (hereafter, 'CUP'), 1910–1913. 666 + 742 + 491 pages. Print run: 750, 500, 500 copies.

Manuscripts. Main manuscript destroyed, but some rejected folios, a concordance index and relevant correspondence held in the Bertrand Russell Archives, McMaster University, Canada. No *Nachlass* for Whitehead.

Later editions. 2nd edition, 3 volumes, 1925–1927, CUP. 674 + 772 + 491 pages. Various photoreprints until the 1960s, then 1997. Pirate photoreprint: Taipei: Rainbow Bridge Publishing Company, 1955(?).

*Abridged version. Principia mathematica to *62*, 1962, CUP.

Part German translation. Einführung in die mathematischen Logik (trans. H. Mokre), Munich and Berlin: Drei Masken, 1932. [Repr. Vienna: Medusa, 1984. The introductory material of the two editions.]

Related articles: Cantor (§46), Dedekind and Peano (§47), Gödel (§71), Hilbert and Bernays (§77).

1 THE REDUCTIONIST HERITAGE

Principia mathematica (hereafter, '*PM*') was the cumulation of nearly a decade of work by its authors that brought to a climax some decades of a search both for rigour in mathemat-

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

ical proof and examination of the relationship between mathematical logic and the closest attendant branches of mathematics. A.N. Whitehead (1861–1947) and Bertrand Russell (1872–1970) argued in this book for a position that, following the philosopher Rudolf Carnap (1891–1970) in the late 1920s, is called 'logicism'; namely, that their logic, together with set theory, supplied not only all the means of deduction required in mathematics but also its objects, starting out from definitions of cardinal and ordinal integers in terms of 'classes' (to quote the technical term then commonly used).

Russell set most of the main guidelines for logicism. He had graduated from Cambridge University in 1894 in mathematics and philosophy, and for the rest of the decade he united the two disciplines in a search for the reasons why mathematical knowledge was secure. Starting off with *An essay on the foundations of geometry* (1897), he then moved to arithmetic and some aspects of mathematical analysis. He was working within neo-Hegelianism, the prevailing Cambridge philosophy of a strongly idealist character which asserted that all be in the mind, especially the 'synthesis' of opposing thesis and antithesis. He was not satisfied with any of the systems obtained, and during 1899 he followed his philosopher friend G.E. Moore in replacing that philosophy with an opposite empiricism, which minimised the role of abstract objects in knowledge [Griffin, 1991]. But this transformation left the mathematical basis still wanting—until a magic morning on Friday 3 August 1900 at the First International Congress of Philosophy, when he learnt of a programme in progress at Turin under the direction of Giuseppe Peano (1858–1932).

Although never a student, Peano knew well the lecture courses of Karl Weierstrass (1815–1897) at Berlin University, in which Weierstrass sought to improve the level of rigour in arithmetic and mathematical analysis, refining the aims already set by A.L. Cauchy (§25): special emphasis was laid upon careful definitions, and provision of all details of proof. Peano tried to underpin arithmetic by offering axioms using new primitives from which integers could be defined, but he was only partially successful (§47). He also greatly increased the formalisation of and symbolism for mathematics; and also the 'mathematical logic' (his name) with which a theory was expressed. He formulated the classical two-valued logic based upon the law of excluded middle ('LEM'), that every proposition is either true or is false, and incorporated both the propositional and the predicate calculi with quantification. His programme consisted in rendering both this logic and mathematical theories in axiomatic forms, and expressing as precisely as possible the basic concepts, definitions and proofs of theorems [Borga et alii, 1985; Rodriguez-Consuegra, 1991, ch. 5]. He and his three main followers spent that Paris morning in August 1900 describing aspects of the project: Alessandro Padoa in person, Cesare Burali-Forti and Mario Pieri with read papers.

The Peanists (as they were known) made great use of the set theory of Georg Cantor (1845–1918). This too had grown out of Weierstrass's programme, and provided not only much machinery for mathematical analysis but also an (apparently) natural link to logic via the association between a 'propositional function' (Peano again) such as 'x is an even number' and the class of even numbers, and quantification represented by relationships between classes such as intersection and inclusion.

Cantor and the Peanists were the two principal influences upon Russell. His first reaction was surprise that Peano's logic did not included a logic of relations; so he promptly supplied one in 1900, adapting logical and set-theoretic notions as appropriate—and largely

ignoring the tradition of such a logic among the algebraic logicists, especially C.S. Peirce and Ernst Schröder [Brady, 2000]. (Relations are propositional functions of more than one variable; below 'propositional function' will include relations also.) Then, he noticed that, while the Peanists laid out mathematical and logical notions (with symbols) in separate columns, set-theoretic notions could appear in both of them; so early in 1901 he decided that *no* division existed, and formulated his logicist thesis that all mathematical notions were already logical ones. So at least the vision was complete and clear; everything came down to classes, which themselves were definable within mathematical logic.

But soon after the joy came sorrow; within a few months Russell found a paradox in set theory. Cantor had a proof that the cardinality of any class was less than that of the class of all its sub-classes, and by adapting it to the class of all classes Russell found that the class of all classes that did not belong to themselves belonged to itself *if and only if* it did not do so. The italicised clause, a bi-implication, underlies the seriousness of the result; in a system based upon logic of all subjects, the foundations were themselves susceptible to contradiction. The apparently natural association of propositional function and class, or the LEM, was lost.

By the time of this discovery Russell was well into writing a large book outlining this position, which appeared in 1903 from the Cambridge University Press as *The principles of mathematics* [Russell, 1903]. The logicist thesis was stated in terms of conditionals: 'if p then q', where p and q were propositions drawing only upon logical (and set-theoretic) notions. He characterised this logical form as 'pure mathematics'—a non-standard and most misleading use of this name. His paradox were presented, with others including those of the greatest ordinal and cardinal number (each was both equal to and larger than itself), and a tentative solution was offered [Garciadiego, 1992]. He also publicised the work of Gottlob Frege (1848–1925), who had already formulated a version of logicism, restricted to arithmetic and some mathematical analysis; Russell had not read his work until after completing the bulk of his book. Frege's logical system was also susceptible to Russell's paradox.

2 COLLABORATION, AND FALLOW YEARS

What to do now? A great vision was in place, but the ground floor had collapsed, and the tentative solution did not work.

Also present in Paris in 1900 had been Whitehead, formerly a tutor of Russell and author of a recent survey of several modern algebras, especially Hermann Grassmann's (§32), in his (misnamed) book *Universal algebra* (1898). Also impressed by Peano and drawn to Cantor, Whitehead had taken up transfinite arithmetic; but around 1904 he joined Russell in the aim of expounding logicism in full symbolic Peanist detail while also Solving those damned paradoxes. Throughout the following years Whitehead was still at Cambridge, but in 1905 Russell had a house built near Oxford, and much of his research was done there. Following a bad practice, I shall refer just to 'Russell' below, though Whitehead was often also involved in the moves [Lowe, 1985].

The paradoxes dominated Russell's efforts for some years. He tried many possibilities, some based upon analysing Cantor's proof method for the power-class theorem, but without success; either some version of the paradox would recur, and/or else some part of the

mathematics was lost. But one important advance was finding in 1905 the means of expressing mathematical functions in terms of propositional ones, with his theory of definite descriptions. Inspired by the need for mathematical functions to be single-valued ($+\sqrt{x}$ is a function but $\sqrt{x}$ is not) he offered conditions for this property to be satisfied contextually within a proposition: at least and at most one object fitted the description involved, and moreover had (or did not have) some property asserted in that proposition. These conditions had already been suggested by Peano for mathematics; Russell saw their bearing on language more generally. His insight was that, in connection with negation, 'the' had to be handled like a quantifier; just as the true proposition 'not all numbers are even' is distinguished from the false 'all numbers are not even', so the negation of 'the present King of France is bald' is not 'the present King of France is not bald', both of which are false, but 'it is not the case that the present King of France is bald', which is true (as also is 'it is ... not bald'). Thereby the LEM was preserved. Mathematical expressions were to be construed this way: for example, the (false) proposition $2 + 3 = 56$ becomes 'the value of $2 + 3$ is 56'.

Thus treatment of descriptions soon led Russell to a 'substitutional' theory, which assumed only propositions and their truth-values and individuals; but for wares reasons, including a form of his paradox, it was abandoned. Retaining the theory of descriptions, from 1907 Russell went back to propositional functions with normal variables and quantification, and he and Whitehead worked out their logicist system. Parts of the developing theory were written up by Russell as papers in the *American journal of mathematics,* with more philosophical aspects being rehearsed in *Mind* and elsewhere.

3 THE WRITING AND CONTENT OF *PM*

The project was divided up into Parts, with one author taking the initial responsibility for its content, which would be checked and criticised by the other, and possibly back again. Russell wrote out the final version for printing; the manuscript (which they destroyed soon after publication) seems to have contained at least 6000 folios, with each theorem and proof written on a separate folio. By the autumn of 1909 the manuscript of the first three volumes was ready for Cambridge University Press; funding was obtained from the Royal Society of London to support printing costs. The title 'Principia mathematica' had been chosen around 1906; maybe it alluded to their Trinity College predecessor Newton (§5), but another candidate was their empiricist leader Moore, whose *Principia ethica* had appeared in 1903. Table 1 lists the logical/mathematical topics covered, and shows that Peano and Cantor dominated [Grattan-Guinness, 2000, ch. 7]; the authors did not even reach the calculus, although it had been treated, in a prosodic manner, in Russell's *The principles.*

The opening Part dealt with the propositional and predicate calculi, including both individual and functional (and relational) quantification. It is the best known and least clear portion of *PM*. The difficulty concerns incoherence of expression; it largely sprang from the inherited Peanist belief that logic was an absolutely general discipline, so that (as we now say) there was no room to talk *about* it. For example, the LEM, which is (and was) normally construed as a metalogical principle, was taken instead to be the proposition

Table 1. Summary by Sections of *Principia mathematica* (1910–1913).
The titles of the Parts, and numbers of pages (omitting the introductions) were: I. 'Mathematical logic' (251); II. 'Prolegomena to cardinal arithmetic' (322); III. 'Cardinal arithmetic' (296); IV. 'Relation-arithmetic' (210); V. 'Series' (490); VI. 'Quantity' (257). '+' indicates surviving pertinent manuscripts.

Section; pages	**(Short) 'Title' or Description: other included topics**
IA: *1–*5; 41	'Theory of deduction': Propositional calculus, axioms.
IB: *9–*14; 65	'Theory of apparent variables': Predicate calculus, types, identity, definite descriptions.
IC: *20–*25, +; 48	'Classes and relations': Basic calculi: empty, non-empty and universal.
ID: *30–*38, +; 73	'Logic of relations': Referents and relata, Converse(s).
IE: *40–*43; 26	'Products and sums of classes': Relative product.
IIA: *50–*56; 57	'Unit classes and couples': Diversity; cardinal 1 and ordinal 2.
IIB: *60–*65; 33	'Sub-classes' and 'sub-relations': Membership, marking types.
IIC: *70–*73; 63	'One-many, many-one, many-many relations': Similarity of classes.
IID: *80–*88, +; 69	'Selections': Multiplicative axiom, existence of its class.
IIE: *90–*97; 98	'Inductive relations': Ancestral, fields, 'posterity of a term'.
IIIA: *100–*106; 63	'Definitions of cardinal numbers': Finite arithmetic, assignment to types.
IIIB: *110–*117; 121	'Addition, multiplication and exponentiation' of finite cardinals: inequalities.
IIIC: *118–*126; 112	'Finite and infinite': Inductive and reflexive cardinals, $\aleph_0$, axiom of infinity.
IVA: *150–*155, +; 46	'Ordinal similarity': Small 'relation-numbers' assigned to types.
IVB: *160–*166; 56	'Addition' and 'product' of relations: Adding a term to a relation, likeness.
IVC: *170–*177; 71	'Multiplication and exponentiation of relations': Relations between sub-classes, laws of relation-arithmetic.
IVD: *180–*186; 38	'Arithmetic of relation-numbers': Addition, products and powers.
VA: *200–*208, +; 97	'General theory of series': Generating relations, 'correlation of series'.
VB: *210–*217; 103	'Sections, segments, stretches': Derived series, Dedekind continuity.
VC: *230–*234; 58	'Convergence' and 'limits of functions': Continuity, oscillation.
VD: *250–*259,+; 107	'Well-ordered series': Ordinals', their inequalities, well-ordering theorem.

Table 1. (*Continued*)

Section; pages	**(Short) 'Title' or Description: other included topics**
VE: *260–*265, +; 71	'Finite and infinite series and ordinals': 'Progressions', 'series of alephs'.
VF: *270–*276; 52	Compact, rational and continuous series: Properties of sub-series.
VIA: *300–*314; 105	'Generalisation of number': Negative integers, ratios and real numbers.
VIB: *330–*337; 58	'Vector-families': 'Open families', vectors as directed magnitudes.
VIC: *350–*359; 50	'Measurement': Coordinates, real numbers as measures.
VID: *370–*375; 35	'Cyclic families': Non-open families, such as angles.

'$p \vee$ not-p' *in* the calculus. Substitutes for the distinction were made for certain circumstances, for example, taken from Frege, p merely considered and $\vdash p$ when asserted with a truth-value.

The most serious casualty of these conflations was implication ('if ... then'), which became muddled with inference, entailment and logical consequence. An example is logicism itself: instead of the conditional form '$\vdash p \supset q$' of *The principles*, they seems to have had in mind its inferential cousin '$\vdash p \supset \vdash q$', but they never made it clear. Among other conflations, that between a (quantifiable) variable and a schematic letter was not made, nor was a rule of substitution offered; and both axioms and rules of inference were bundled together under the Peanist title 'primitive propositions'.

The paradoxes were avoided (Solved?) by the 'vicious circle principle', which stated that any object defined in terms of a class of objects cannot belong to that class. Thus an object could belonged only to a class which was one 'type' up, classes only to classes of classes, and so on, including for classes of ordered pairs, ordered trios, and so on. A class was defined from an appropriate propositional function contextually in a proposition, and the 'order' was specified by determining which variables were quantified within that function. Self-membership of any class was avoided, thereby stopping paradoxes from being formed. In addition, there was a hierarchy of types of propositions, so that propositional paradoxes, such as the ancient one involving 'this proposition is false', could also be Solved: 'this' was no longer self-referential, and the truth-value of the proposition lay in the type above that of the proposition to which it referred.

The mathematical treatment was based upon defining cardinals and ordinals as classes of equipollent classes and of well-order series respectively, starting out with their zeroes defined respectively as the class of the empty class and of the empty relation. However, type theory caused mathematical objects to be sited in different places; for example, the numbers 2, $5/78$ and $\sqrt{13}$ were in different types, so that they could not be handled together arithmetically. So Russell reluctantly proposed an 'axiom of reducibility', which assumed that for any propositional function there existed a logically equivalent one free from quantifiers; arithmetical operations were thereby restored, but type theory was truncated.

Further, the reconstruction of Cantor's transfinite arithmetic needed an axiom of infinity, which required that the bottom type be composed of an infinitude of individuals. Even then, the construction of Cantor's theory was only partial; for only a finite number of types was allowed, so that the numbers ω_ω and $\aleph_\omega$ and beyond could not be defined anyway. But Moore's empiricism had interfered already: these individuals, basic structureless objects, could not be abstract, and they were hardly logical, so they had to be physical. Thus the axiom was an empirical proposition; but then mathematical logic became *a posteriori.* To minimise this defect, Russell stipulated that when the full roster of individuals was not required, only one of them was to be assumed; however, Whitehead had forgotten this rule when preparing the exegesis of cardinal arithmetic for volume 2 of the book, and several sections had to be rewritten on proof during 1911, with a new preface written by Whitehead.

A final difficulty was the axiom of choice, to use the name given to it by Ernst Zermelo after he introduced it in 1904 to prove Cantor's well-ordering theorem. Slightly earlier that year Russell had himself recognised the need for this axiom, when trying to define the multiplication of an infinitude of numbers in terms of classes of classes, and called it the 'multiplicative axiom'. There was a strong debate about these axioms; their various forms, the places for their need, and the philosophical legitimacy of its non-constructive character [Moore, 1982]. *PM* includes a fine account of its forms and role as understood around 1910, for it posed an extra philosophical worry for Russell: the mathematical logic in *PM* was *finitistic*, in both the length of propositions and of proofs, but this axiom allowed the execution of an *infinitude* of independent operations, so that its expression within its logic was very problematic.

But there was much to please as well as to worry in *PM*. The use of relations was quite virtuoso; in particular, the logic of relations was deployed widely to express all sorts of properties, often starting out from that between the argument and the value of a mathematical function. A high point occurs in volume 2 with a superb generalisation of ordinal arithmetic, which Russell had conceived soon after learning Peano's system: relation-arithmetic', a calculus of "numbers" defined as classes of 'ordinally similar' relations of which the orthodox arithmetic of ordinals based upon well-ordering relations was a special case.

Missing was a treatment of geometry, which Whitehead promised to write alone as volume 4 [Harrell, 1988]; much of the material in the last sections of volume 3 had been prepared for its benefit. Aspects of projective, metrical and descriptive branches were to be treated, together with a 'construction of space' based upon the logic of relations that he had already outlined in a paper of 1906 [Grattan-Guinness, 2002]; apparently three- and four-place relations would be deployed extensively. He seems to have written quite lot of this volume; but he gave it up around 1918, and the manuscripts were destroyed after his death in 1947. But even had it been finished, some significant geometry would have still been missing; for example, differential geometry.

4 REACTIONS BY RUSSELL AND HIS BRITISH FOLLOWERS

PM was quite widely read and used, at least in parts [Grattan-Guinness, 2000, chs. 8–9]. The audience included not only logicians (a small community at that time) but also

philosophers (normally the logic, sometimes also the logicism) and some mathematicians (among whom the interest was rather slight, so that, for example, the importance of *PM* in the development of set theory is little recognised). On logicism the three troublesome axioms gained much attention, the limited coverage of mathematics rather little.

After the appearance of *PM*, apart from Whitehead's volume 4 both he and Russell largely abandoned logic(ism) for philosophy, of very different kinds. Russell maintained the same Moorean empiricism in his epistemology; and logical techniques, especially those involving relations, played a prominent role. *Our knowledge of the external world as a field for scientific method in philosophy* (1914) was very influential. In 1919 he published a popular *Introduction to mathematical philosophy,* which treated the main features of *PM* (geometry was avoided): definition of integers, order relations and ordinals, Cantorian transfinite arithmetic, limits and continuity, the axioms of choice and of infinity, type theory and the paradoxes (where he confessed that assuming the existence of at least one individual was 'a defect in logical purity': [Russell, 1919, 203]). By contrast, the account of logic itself was rather cursory.

In that year Russell also resumed his friendship with his pre-War student Ludwig Wittgenstein (1889–1951), who in the interim had written a philosophical work which was to appear with an introduction by Russell as *Tractatus logico-philosophicus* (1921, 1922). While not concerned with logicism (he had nothing to say about infinite sets, for example), Wittgenstein had realised that Russell's logic and logicism had become mixed together. So he tried to characterise logic independently, in an extensional manner with the connectives construed as functorial compounds of truth-values and logical propositions defined as tautologies or contradictions. In his introduction Russell reacted against the ensuing monism and envisioned a hierarchy of languages, each of which could talk about those below as well as about the world. It was one of his greatest philosophical suggestions; yet he never appreciated its significance. Only two years later he prepared new material for a second edition of *PM,* a repeat of the first one with proposed revisions, but hierarchies of theory *played no role.* The changes included some use of truth functions, although the notion of tautology was not used.

For some reason Cambridge University Press reset the first two volumes (unchanged?); Russell was helped in their reading by Frank Ramsey (1903–1930), who then outlined his own version of logicism that was even more extensional than Russell's revision. He even defined the universal and existential quantifiers as infinite conjunctions and disjunctions respectively, with no concern over the horizontally infinitary logic that could be involved. He used the notion of tautology to propose a way in which the axioms of reducibility and of infinity could be rendered logical (in his sense). He also pointed out that some paradoxes were concerned with naming as such rather than with mathematics, and followers of this distinction have regarded them as irrelevant to logicism.

After Ramsey's early death *PM* did not gain much interest in Britain, though Susan Stebbing (1885–1943) flew his flag against the continuing opposition of philosophers at Oxford and Cambridge. Russell himself revived his academic life in the mid 1930s, and reprinted *The principles* in 1937 with a new preface; however, and regrettably not for the only time, he misstated logicism as an *identity* thesis between mathematics and logic! The centre of gravity of Anglo-Saxon attention had long shifted across the Atlantic.

5 THE RECEPTION OF *PM* IN THE UNITED STATES

Russell's logical programme had aroused the curiosity of some American mathematicians already in the 1900s. After its publication *PM* stimulated much interest, especially with a two-month visit made by Russell in the spring of 1914 to Harvard University where he gave two lecture courses, one working though *PM* and the other heralding *Our knowledge.* (Extensive notes were taken by a graduate student called T.S. Eliot.) One important reaction was negative: the recent graduate C.I. Lewis (1883–1964) noted the confusions in the various uses of implication and proposed to distinguish the material from its 'strict' versions. Thereby he became the main founder of modal logics.

Lewis's sensitivity to this issue was typical of interested Americans. One reason was the development there of model theory [Scanlan, 1991], where (meta) properties of formal systems was a main focus. Over the years E.V. Huntington (1874–1952) included parts of *PM* among his bestiary of mathematical theories; and H.M. Sheffer (1882–1964), a fellow student of Lewis at Harvard, was concerned with features of formal systems (such as introducing the neither-nor connective now named after him, in a paper of 1913 which also inaugurated the name 'Boolean algebras'). In a review of 1926 of the second edition of *PM* Sheffer stated this conundrum more generally than anyone else: '*In order to give an account of logic, we must presuppose and employ logic*', and he saw no way out of this 'logocentric predicament' [Sheffer, 1926, 228]. Yet despite the American sensitivity to this issue, the breakthrough to full metalogic was to lie in European hands.

6 GERMAN-SPEAKING CONTRIBUTIONS

When Russell met Peano in 1900 David Hilbert (1862–1943) was beginning his own studies of the foundations of mathematics, after his work on geometry (§55). The fruits were not great, and he largely stopped in 1905; but a second phase began in 1917, with the 'metamathematical' study of foundations, set theory the articulation of 'proof theory' that was to flower into the 1930s (§77). *PM* had a welcome place at first; from 1918 it furnished the axioms of the calculi, and until reflecting upon the three troublesome axioms Hilbert thought that its reduction of mathematics to that logic was successful. But then the differences of approach became clearer, and the acrimonious disagreement of the late 1920s over foundations between Hilbert and the intuitionist L.E.J. Brouwer (1881–1966), who introduced the insulting name 'formalism' for Hilbert's stance that unfortunately has become standard, came to eclipse logicism as a general philosophy of mathematics. However, the logic of *PM* had secured a good place, and maintained it. For philosophers the situation was less warm; several of a neo-Kantian or phenomenological hue, such as Ernst Cassirer and Gerhard Stammler, doubted that mathematics was no longer synthetic *a priori*. There was also some posthumous appreciation of Frege's contribution to logic and their ramifications for the philosophy of language (more than of his logicism, in fact).

The other main German-speaking centre of attention lay with the Vienna Circle of philosophers, where Carnap was especially significant. From the mid 1920s he not only

formulated Russell's version of logicism but also tried to formalise the epistemology of *Our knowledge* in a way which Russell had never attempted; the main outcomes were two books, the forgotten *Abriss der Logistik mit besondere Berücksichtigung der Relationstheorie und ihre Anwendungen* (1929), where the word 'logicism' was proposed ('logistic' had been widely used since the mid 1900s, but to refer both to Peano's and Russell's positions); and the famous *Der logische Aufbau der Welt* (1928).

But the greatest impact upon logicism came from a young graduate of the University of Vienna, Kurt Gödel (1906–1978), who in 1931 proved his famous theorem that any axiomatised theory encompassing number theory with quantification over integers could never be fully axiomatised, and that the establishment of its consistency needed a richer theory. This paper is the subject of §71 below; we note here that the theorems refuted hopes for completeness and consistency that had been stated in the introduction of *PM*, and also that *the central importance of the distinction between logic and metalogic* was clarified in the paper.

Gödel was a founder of the importance of this distinction, doubtless partly guided by its presence in Hilbert's proof theory. Its other main father-figure was the Pole Alfred Tarski (1901–1983), who came to it by a somewhat different route much less linked to logicism or *PM*. But Russell's paradox had also played a role in the work of Jan Łukasiewicz (1878–1956) and Stanisław Leśniewski (1886–1939), the leaders of a strong development of formal logic in Poland from the late 1910s [Wolenski, 1989]. Łukasiewicz was led partly by the paradox to consider three- and many-valued logics (not to be confused with Lewis's modal logics), while Leśniewski started out from two different readings of the paradoxical argument to generate two large-scale formal systems; a largely extensional 'Mereology' and an intensional 'Ontology'. Outside this circle, in the 1920s Leon Chwistek (1884–1944) anticipated Ramsey on the two kinds of paradox, and also tried to rebuild *PM* while drawing also upon aspects of metamathematics.

7 LOGIC(ISM) AFTER GÖDEL

After the absorption of Gödel's theorem, the meta-questions posed about mathematical and logical systems had to be changed. From many later works, those of the Harvard logician W.V. Quine (1908–2000), nominally a student of Whitehead at Harvard in the early 1930s, make a fitting conclusion. He too set up large-scale logical systems using set theory; but in the post-Gödelian atmosphere he did not try to reduce everything solely to logical notions, and indeed became a major actor in exploring the plurality of relationships available [Quine, 1969].

The aims of Whitehead and Russell had been refuted by Gödel, and questioned from other points of view earlier; but many valuable technical procedures and philosophical issues had been profitably examined. For example, Gödel might not have envisioned his theorem on the limitations of such endeavours. And the irony of 1922 continued; for even into his nineties Russell was still trying, and failing, to understand the significance of Gödel's theorem [Grattan-Guinness, 2000, 592–593].

BIBLIOGRAPHY

Borga, M., Freguglia, P and Palladino, D. 1985. *I contributi fondazionale della scuola di Peano*, Milan: Franco Angeli.

Brady, G. 2000. *From Peirce to Skolem. A neglected chapter in the history of logic*, Amsterdam: Elsevier.

Garciadiego, A. 1992. *Bertrand Russell and the origins of the set-theoretic 'paradoxes'*, Basel: Birkhäuser.

Grattan-Guinness, I. 2000. *The search for mathematical roots, 1870–1940. Logics, set theories and the foundations of mathematics from Cantor through Russell to Gödel*, Princeton: Princeton University Press.

Grattan-Guinness, I. 2002. 'Algebras, projective geometry, mathematical logic, and constructing the world: intersections in the philosophy of mathematics of A.N. Whitehead', *Historia mathematica*, *29*, 427–462. [Printing correction: *30* (2003), 96.]

Griffin, N. 1991. *Russell's idealist apprenticeship*, Oxford: Clarendon Press.

Harrell, M. 1988. 'Extension to geometry of *PM* and related systems II', *Russell, n.s. 8*, 140–160.

Lowe, V. 1985. *Alfred North Whitehead. The man and his work*, vol. 1, Baltimore: Johns Hopkins University Press.

Moore, G.H. 1982. *Zermelo's axiom of choice*, New York: Springer.

Quine, W.V.O. 1969. *Set theory and its logic*, 2nd ed., Cambridge, MA: Harvard University Press.

Rodriguez-Consuegra, F.A. 1991. *The mathematical philosophy of Bertrand Russell: origins and development*, Basel: Birkhäuser.

Russell, B. 1903. *The principles of mathematics*, Cambridge: Cambridge University Press. [Repr. with new introd. London: Allen & Unwin, 1937.]

Russell, B. 1919. *Introduction to mathematical philosophy*, London: Allen & Unwin; New York: Macmillan. [Various reprs. and transs.]

Russell, B. *Collected papers.* About 30 vols., London: [now] Routledge, 1983. [See especially vols. 1–9 (vol. 5 still not ready).]

Russell, B. 1967. *The autobiography*, vol. 1, London: Allen & Unwin.

Russell: the journal of the Russell Archives, 1971–, McMaster University Press.

Scanlan, M.J. 1991. 'Who were the American postulate theorists?', *Journal of symbolic logic*, *56*, 981–1002.

Sheffer, H.M. 1926. Review of *PM*, 2nd ed., vol. 1, *Isis*, *8*, 226–231.

Wolenski, J. 1989. *Logic and philosophy in the Lvov–Warsaw school*, Dordrecht: Kluwer.

CHAPTER 62

FEDERIGO ENRIQUES AND OSCAR CHISINI, LECTURES ON 'THE GEOMETRICAL THEORY OF EQUATIONS AND ALGEBRAIC FUNCTIONS' (1915–1934)

A. Conte

In these extensive volumes Enriques and Chisini gave a detailed and authoritative account of algebraic geometry. Thanks especially to Enriques's approach to mathematics, they included much historical information about the subject. They built much upon the Italian tradition in it; but a polemic with Francesco Severi was to develop, mainly over questions concerning rigour.

First publication. Lezioni sulla teoria geometrica delle equazioni e delle funzioni algebriche, 4 volumes, Bologna: Zanichelli, 1915, 1918, 1924, 1934. xiv + 398, 713, 594, viii + 274 pages. [Vols. 1–3 have only Enriques as author, and are 'edited by Dr. Oscar Chisini'; vol. 4 carries both names as authors.]

Photoreprint. 2 volumes, Bologna: Zanichelli, 1985.

Related articles: Riemann (§34, §39).

1 THE AUTHORS

Federigo Enriques, together with Guido Castelnuovo (1865–1952) and Francesco Severi (1879–1961), make up a famous triad of great masters who shaped the Italian school of algebraic geometry in the 20th century. Enriques was born in Livorno in 1871 and received his degree at the *Scuola Normale Superiore* of Pisa with Eugenio Bertini in 1891. After remaining in 1892 in Pisa, he moved in 1893 to Rome, where he started his scientific collaboration with Castelnuovo; and in 1894 to Turin, in order to study with Corrado Segre. That same year, after becoming free docent, he started to teach at the University

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

of Bologna; in 1897 he won the competition for the chair of Projective and Descriptive Geometry, and he stayed until 1923. Then he moved on the chair of Higher Geometry at the University of Rome, joining both Castelnuovo and Severi, who had been professors there since 1891 and 1922 respectively.

Enriques was an eclectic personality, whose interests were not confined to algebraic geometry but took in also the foundations, psychology and philosophy of mathematics together with the history and philosophy of science. His main contribution to algebraic geometry was the classification of algebraic surfaces, probably the most outstanding legacy of the Italian school of algebraic geometry, which he achieved in a long series of papers running from 1896 to 1914, some of them in collaboration with Castelnuovo, who contributed too some of its essential tools, such as the proof of the Riemann–Roch theorem for surfaces and his celebrated criterion of rationality. The genesis of the classification can be traced in the letters of Enriques to Castelnuovo [Enriques, 1996], whilst its final version is to be found in [Enriques and Campedelli, 1932] and in the posthumous volume [Enriques, 1949] edited by Castelnuovo.

Enriques had a fascinating personality, which attracted the best students to build a large and important school, whose more important members were, besides Chisini, Luigi Campedelli (1903–1978), Fabio Conforto (1909–1954), Giuseppe Pompily (1913–1968), Alfredo Franchetta (1916–) and the Belgian Lucien Godeaux (1887–1975). In a broader sense can be counted among Enriques's students also the Pole Oscar Zariski (1899–1986); he spent the years 1921–1927 in Rome, where he was deeply influenced by the type of problems that Enriques was studying (and at Enriques's suggestion changed his name from the original 'Ascher Zaritski'), even if he followed eventually Castelnuovo's vision of algebraic geometry. Afterwards Zariski moved to Harvard University, where he created one of the most important schools of algebraic geometry of the last century.

In 1938 Enriques fell victim, like all Jewish professors, of the racial laws of the fascist régime and was compelled to quit the University. He was not even allowed to put his name on [Enriques and Conforto, 1939], which appeared under the only name of Conforto. He died in Rome in 1946.

Oscar Chisini was born in 1889 in Bergamo, near Milan, as the son of an army officer. The career of the father brought the family to Bologna at the end of the century, and it was in this town's very ancient University that he received his degree in 1912. Having showed a very bright intellect, Enriques asked him to collaborate in the drafting of the present *Lezioni* even though he was still very young and in the first stages of the study of algebraic geometry. This collaboration was to last more than twenty years. Chisini's career started as free docent in 1918; then he won the competition for a chair in 1923, and taught first at Cagliari and then in Milan, where he remained until his retirement in 1959, and died in 1967. He was profoundly influenced by the collaboration, and most of his research interests started out from topics covered by it: among others, the theory of singularities both of curves and surfaces, and the related questions of degeneration, with ingenious applications to branch curves of multiple planes and their 'limit forms'; the realization of a visible model of the fundamental group of the complement of an algebraic curve in the complex projective plane, called by him 'characteristic braid' ('treccia caratteristica') of the algebraic curve, and related to the 'braid group' studied also in the 1930s by Emil Artin and his school; and the topological approach to topics such as the

theory of correspondences and that of intersection multiplicities between two algebraic curves.

2 THE LECTURES

The *Lezioni* are noteworthy from many points of view: first of all, for their monumental size (more than 2000 pages); then for being a true work of collaboration, following Enriques's habit of involving his best students in the editing of his treatises—as he will do with Campedelli for [Enriques and Campedelli, 1932] and with Conforto for [Enriques and Conforto, 1939]; and finally for the peculiar way in which they were created [Manara, 1968]:

> Chisini related that the work itself had been conceived almost completely in a 'peripatetic' fashion , as he used to say: that is, by strolling under the porticoes of Bologna with Enriques; even the most complicated formal arguments [. . .] had not been conceived at a desk; at the most, Enriques stopped and wrote with the tip of his umbrella on the pavement of Bologna's porticoes, while they were strolling along.

Their four volumes, published between 1915 to 1934, built up a complete treatise on the theory of algebraic curves, which are studied from all possible points of view: synthetic-projective, analytic-differential and topological-transcendental. Moreover, they not only collected all the principal results of the theory, but gave also a massive amount of specific examples, treated in all details, and lengthy historical notes, tracing the origins of the various parts of the theory. To grasp their content, the best way is to follow the advice of the authors (vol. 1, xiii–xiv) and read the titles of the Books and Chapters. These are given in Table 1.

3 THE METHODS AND THE CONTENT

The *Lezioni* are written in the classical style of the Italian school of algebraic geometry going back to Luigi Cremona (1830–1903), depending on the methods of projective geometry and on that unrivalled knowledge of the behaviour of algebraic curves and surfaces sitting in projective spaces which was peculiar of the great masters of the school, and unsurpassed in Enriques. Moreover, besides the 'geometric algebra' of Cremona, Corrado Segre and Castelnuovo, the innovative element in Enriques and Chisini's approach is the importance that they attribute, as already remarked, to the topological-transcendental vision of the theory of algebraic curves.

As to the method followed by the authors, probably its most charming aspect is the heuristic approach, which consists essentially in the 'exhibition, next to the truths, of the paths—often different—that led there, without excluding from the comparison of methods any partial or imperfect procedures, and rather with the precise intention of letting them correct and clarify themselves mutually, making clear how much is missing in each partial conception of the theories' (vol. 1, x). From this premise it follows naturally the necessity of a strong historical approach in order, as Enriques will repeat in his last work [Enriques, 1949, x],

Table 1. Summary by Books and Chapters of Enriques's and Chisini's lectures. The four volumes divide thus: Books I–II; III–IV; V; and VI.

Bk., ch.; pp.	'**Title**'
I; 148	'Introduction'.
I,I; 50	'The equations $f(x) = 0$ and the groups of points on the line'.
I,II; 72	'The fundamental interpretations of the equation $f(x) = 0$: curves and correspondences'.
I,III; 26	'Note on the meaning of the expression "in general" and on the computation of constants'.
II; 239	'The principle of correspondence and its applications'.
II,I; 64	'The involutions and the finite groups of projectivities on the line'.
II,II; 122	'Elementary theory of plane curves'.
II,III; 53	'Note on algebraic functions and on the real representations of the imaginary'.
III; 306	'The elementary theory of plane curves based on polarity'.
III,I; 72	'Polarity and covariant curves'.
III,II; 112	'The problem of intersections and the Plückerian characters of curves'.
III,III; 42	'The plane cubic'.
III,IV; 80	'Appendix: reality and continuity; enumerative geometry'.
IV; 359	'The singularities of algebraic curves'.
IV,I; 74	'Singularities and Puiseux series developments'.
IV,II; 58	'Singularities with respect to quadratic transformations'.
IV,III; 86	'Singularities with respect to the differential calculus'.
IV,IV; 141	'Appendix: singularities of skew space curves and of surfaces'.
V; 568	'Curves and algebraic functions of one variable'.
V,I; 120	'Linear series on a curve'.
V,II; 80	'The geometry of plane curves and Cremona transformations: historical evolution of ideas'.
V,III; 224	'Curves and transformations'.
V,IV; 84	'Correspondences between curves'.
V,V; 60	'On the theory of skew space curves'.
VI; 264	'Elliptic and abelian functions'.
VI,I; 104	'Elliptic integrals and functions'.
VI,II; 46	'Abelian integrals'.
VI,III; 114	'The problem of inversion and Abelian functions'.

> not only to give to the expounded theories a logical structure, but also [...] to give an historical perspective of their coming into being. In this way one wants to offer to the reader, not just the gift of something perfect that one is allowed to look at from the outside, but rather the vision of an acquisition and an advancement, the reasons for which one must understand and which the reader is invited to re-learn by himself and for himself, finding in the book a working tool.

The importance of the many examples and particular cases discussed then follows; often it lies upon their historical significance, which is one of the reasons why the *Lezioni* are still a valuable source for scholars in the theory and classification of algebraic curves. For example, one can look into volume II, book III, chapter III on the plane cubic, or into volume III, book V, chapter V on skew space curves. Each chapter gives all existing information on its topic(s), expounded with clear and complete geometric arguments. From this point of view the question of rigour (the Achilles's heel of the Italian school) should be viewed in a new light; for Enriques 'historical errors, paradoxes and sophisms acquire special interest, in that they often pointed out the way to more important discoveries' (vol. 1, x). The same criterion is applied to the question of the generality of the results (vol. 1, x–xi):

> The research criterion so splendidly taken advantage of by Abel—to pose problems in their most general aspect in order to discover their true nature—set the course of Analysis, which wanted to liberate the knowledge of qualitative relationships from the accidental complications of calculations; this is precisely that course of maximum realization for the geometric theory of equations and algebraic functions. But the precept of generality has received another interpretation among contemporary geometers [...] The maxim has been established that every theorem must always be enunciated in the most general form to which it is susceptible [...] It is proper to recognize that this habit has diminished the powerful effectiveness of the finest masters, and deserves to be seriously opposed. Since, in the first place, an overly abstract statement succeeds in obscuring the true significance of the theorem, hiding its origins, and—in the second place—it charms young scholars with easy, purely formal, generalizations. [...] Every problem has in a way its own proper degree of generality, and that degree is the first in which the problem itself reveals its true nature [...].

In order to give an example of a typical theorem of the *Lezioni* and of its wording, take a famous theorem of Enriques himself on the quadrics going through the canonical curve associated to an algebraic curve (vol. III, 106):

THEOREM. *A curve of genus $p > 4$, non hyperelliptic, has as canonical curve a C^{p-1}_{2p-2} belonging to a linear system of dimension $(p-1)(p-2)-(3p-3)$ of quadrics Q of the projective space S_{p-1}, and is in general defined as the base curve of this system. The only exceptions are when it contains a g^1_3, and then the C^{p-1}_{2p-2} lies on a rational ruled normal scroll of order $p-2$ common to all Q's; and—for $p=6$—also the case when the C^5_{10} contains a g^2_5 and lies therefore on a Veronese surface, through which go all the Q's.*

The meaning of the theorem is that a canonical curve C of genus p and degree $(2p-2)$ sitting in the projective space $\mathbf{P}^{p-1}$ of dimension $(p-1)$ is always the set-theoretical intersection of the quadrics (= hypersurfaces of degree 2) going through it, unless C is trigonal (= contains a g^1_3) or isomorphic to a plane quintic (= contains a g^2_5), and in these two exceptional cases the quadrics through C set-theoretically intersect respectively in a rational

normal ruled scroll of degree $p-2$ or a Veronese surface (= the surface of $\mathbf{P}^5$ representing the linear system of conics of $\mathbf{P}^2$) containing C. The proof is based on a classically beautiful geometric argument which cleverly uses the classification of rational curves and surfaces in projective spaces, with a final application of the theorem of Riemann–Roch for algebraic curves.

This theorem has aroused much interest up to our times for its great importance in the theory of special divisors on algebraic curves. For the subsequent results of Petri in 1922 and 1925, Babbage in 1939, and Saint-Donat in 1974, together with an appreciation of its importance and full bibliographical references, see [Arbarello et alii, 1985, Chapter III, art. 3].

4 THE IMPACT

The *Lezioni* was a great success from the start and became quickly the standard reference work, all over the world, for everything connected with algebraic curves. Indeed, nowhere else one could find, explained in all details and supported by a lot of examples and historical notes, topics such as the correspondence principle, the theory of intersections, the theory of singularities, and the full theory of linear series, with its many applications. In the meantime, before and after the publication of the third volume of the *Lezioni*, Severi published his own two treatises which deal mostly with algebraic curves. The first one [Severi, 1921] is a German translation (due to Eugen Löffler) of a lithographed text produced in Padova in 1908, to which Severi added various appendixes. In particular, Appendices F (devoted to the theory of the moduli of curves) and G (on the classification of non-degenerate space curves), totalling together about 100 pages, are very important for the future developments of the higher theory of algebraic curves. In the second treatise Severi [1926] announced the ambition 'to collect, coordinate and complete, where necessary, everything that is important in the field of algebraic geometry'. However, only the first volume appeared. Concerned with the geometry of linear series on curves, it covered more or less the same topics as volume III of the *Lezioni*, but it aimed to offer a much more rigorous and systematic exposition of the geometry of algebraic curves than that given by Enriques and Chisini.

There was an implicit polemic involved here; and it became explicit in 1934–1935, when Severi criticized as unrigorous the treatment in Chapter III of Book I of the *Lezioni* of the so-called 'Plücker–Clebsch criterion', which stated that a system of r algebraic equations in r unknowns with coefficients depending rationally in various parameters, is in general compatible if it admits a finite number of solutions for a particular set of values of the parameters. In the harsh discussion which followed Severi went so far to raise the accusation that 'in the treatise [...] the proofs of fundamental theorems are only approximate'. Enriques replied sharply that Severi's approach was unnecessarily and excessively abstract and general. For more details on this polemic, see [Brigaglia and Ciliberto, 1995, 64–67].

As a matter of fact, Severi's bold attempt to give sound foundations to algebraic geometry also revealed itself to be unfruitful; one had to wait until after 1945 for the work, based on abstract algebra, of Zariski and his school, and especially of André Weil, Jean-Pierre Serre and Alexander Grothendieck, to find a fully satisfactory solution to the foundational

problems. These attracted the main efforts of researchers in algebraic geometry during the 1950s and 1960s, so that the interest for the classical results of the Italian school diminished almost to zero. However, starting from the beginning of the 1970s, there was a revival, still active, which led in particular to an enormous activity in the theory of algebraic curves, which looks now completely different from the form left by Castelnuovo, Enriques and Severi. In this process the appreciation for the *Lezioni* rose again (including the photoreprint of 1985 listed above), and it is now universally recognized as the masterpiece it is, putting itself as an example to follow even for the most recent and advanced books on the subject. For example, 'These volumes are written in the spirit of the classical treatises on the geometry of curves, such as Enriques–Chisini' [Arbarello et alii, 1985, vii].

BIBLIOGRAPHY

Arbarello, E., Cornalba, M., Griffiths, P.A. and Harris, J. 1985. *Geometry of algebraic curves*, vol. 1, New York: Springer.

Brigaglia, A. and Ciliberto, C. 1995. *Italian algebraic geometry between the two world wars*, Kingston: Queen's University (Queen's Papers in Pure and Applied Mathematics, vol. 100).

Castelnuovo, G. 1947. 'Federigo Enriques', *Rendiconti dell'Accademia Nazionale dei Lincei, (8) 2*, 3–21.

Enriques, F. and Campedelli, L. 1932. *Lezioni sulla teoria delle superficie algebriche*, pt. 1, Padova: Cedam.

Enriques, F. and Conforto, F. 1939. *Le superficie razionali*, Bologna: Zanichelli.

Enriques, F. 1949. *Le superficie algebriche* (ed. G. Castelnuovo), Bologna: Zanichelli.

Enriques, F. 1996. *Riposte armonie. Lettere di Federigo Enriques a Guido Castelnuovo* (ed. U. Bottazzini, A. Conte and P. Gario), Turin: Bollati Boringhieri.

Manara, C.F. 1968. 'Ricordo di Oscar Chisini', *Periodico di matematiche, (4) 46*, 1–20.

Severi, F. 1921. *Vorlesungen über algebraische Geometrie*, Leipzig: Teubner.

Severi, F. 1926. *Trattato di geometria algebrica*, vol. 1, pt. 1, Bologna: Zanichelli.

CHAPTER 63

ALBERT EINSTEIN, REVIEW PAPER ON GENERAL RELATIVITY THEORY (1916)

T. Sauer

This paper was the first comprehensive overview of the final version of Einstein's general theory of relativity after several expositions of preliminary versions and latest revisions of the theory in November 1915. It includes a self-contained exposition of the elements of tensor calculus that are needed for the theory.

First publication. 'Die Grundlage der allgemeinen Relativitätstheorie', *Annalen der Physik, 49* (1916), 769–822. Also published separately, Leipzig: Barth, 1916.

Later editions. Various reprints of the separately printed version, 5th reprint in 1929. Also in the 3rd and later editions of the anthology H.A. Lorentz, A. Einstein and H. Minkowski, *Das Relativitätsprinzip*, Leipzig: Teubner, 1919. Also in *Published writings of Albert Einstein*, Readex Microprint, 1960, item 78. Also in K. von Meyenn (ed.), *Albert Einstein's Relativitätstheorie. Die grundlegenden Arbeiten*, Braunschweig:Vieweg, 1990. First edition repr. with annotations in *Collected papers of Albert Einstein*, vol. 6, Princeton: Princeton University Press, 1996, 283–339 (Doc. 30). German and English reprints of the *Collected papers* version also available online at http://www.alberteinstein.info (2003).

English translations. 1) By S.N. Bose in *The principle of relativity. Original papers by A. Einstein and H. Minkowski*, Calcutta: University of Calcutta Press, 1920, 89–163. 2) By W. Perrett and G.B. Jeffery (without the first page) in H.A. Lorentz et alii, *The principle of relativity*, London: Methuen, 1923 (repr. New York: Dover, 1952), 109–164.

French translations. 1) By M. Solovine in Einstein, *Les fondements de la théorie de la relativité générale. Théorie unitaire de la gravitation et de l'électricité. Sur la structure cosmologique de l'espace*, Paris: Hermann, 1933, 7–71. 2) By F. Balibar et alii in Einstein, *Oeuvres choisies*, vol. 2, Paris: Editions du Seuil, Editions du CNRS, 1993, 179–227.

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

Italian translation. By A. Fratelli in Einstein, *Come io vedo il mondo. La teoria della relatività*. Roma: Newton & Compton, 1988, 114–185.

Russian translation. In Einstein, *Sobranie naychnykh trudov*, vol. 1, Moscow: Izdatel'stvo Nauka, 1965, 452–504.

Spanish translation. By F. Alsina Fuertes and D. Canals Frau, in Albert Einstein, *La relatividad* (*Memorias originales*), Buenos Aires: Emecé éditores, 1950, 115–213.

Manuscripts. A manuscript of 46 pages is in the Schwadron collection at the Hebrew University, Jerusalem; available online at http://www.alberteinstein.info, Call No. 120–788.

Related articles: Newton (§5), Riemann on geometry (§39), Maxwell (§44), Hertz (§52), Kelvin (§58), Lorentz (§60).

1 THE SPECIAL THEORY OF RELATIVITY

Some ten years before the first review of the *general* theory of relativity, Albert Einstein (1879–1955) had published his famous paper 'On the electrodynamics of moving bodies' [Einstein, 1905]. That paper introduced what later became to be called 'the special theory of relativity'. It presented a conceptual analysis of the notions of space and time, with a critical reassessment of the meaning of simultaneity at its core. Its most salient features are length contraction and time dilation in a system that is in uniform relative motion to an observer with a speed comparable to that of light.

The 1905 paper was not a very sophisticated paper on the mathematical side. Its author had obtained a diploma as secondary school teacher for mathematics and physics at the Zurich Polytechnic in 1900 [Pais, 1982; Fölsing, 1998]. His science education had been excellent, with laboratory work in the most up-to-date facilities, and first-rate mathematics teachers such as Adolf Hurwitz (1859–1919), Carl Friedrich Geiser (1843–1934), and Hermann Minkowski (1864–1909). If more recent advances in theoretical physics were somewhat neglected by his physics teacher Heinrich Friedrich Weber (1843–1912), the young Einstein made up for it in extensive autodidactic studies. Fascinated by laboratory experience, he seems to have skipped more than one of his mathematics lectures, though, and obtained his knowledge when preparing for examinations with the help of lecture notes that had been carefully worked out by his more mathematically inclined friend Marcel Grossmann (1878–1936).

After initial attempts to start a traditional academic career had failed, Einstein composed his theory of special relativity in the evening hours after office work as a technical expert, especially for electrotechnology, at the Patent office in Bern. Mathematically, the breakthrough of special relativity came in a representation using only standard techniques of elementary calculus. Maxwell's electromagnetic equations were written component-wise, notwithstanding the fact that compact vector notation had already been well developed, if not standardized, in electrodynamics and hydrodynamics by the end of the 19th century (§35.5).

The subsequent generalization of the special theory of relativity to a generally covariant theory of gravitation proceeded in three major steps, namely 1) the formulation of the

equivalence hypothesis in 1907, 2) the introduction of the *metric tensor* as the crucial mathematical concept for a generally relativistic theory of gravitation in 1912, and 3) the discovery of the generally covariant *field equations of gravitation* in 1915. See [Norton, 1984; Stachel, 1995, and 2002, sec. V; and Renn et alii, forthcoming]; for further references, see the literature cited in these works, and on specific aspects also volumes 1 [Howard and Stachel, 1989], 3 [Eisenstaedt and Kox, 1992], 5 [Earman et alii, 1993], and 7 [Goenner et alii, 1999] of the Einstein Studies series. The *Collected papers of Albert Einstein* are cited hereafter as 'CPAE'.

2 THE EQUIVALENCE HYPOTHESIS

In 1907, Einstein saw himself confronted with the task of reflecting on the consequences of the relativity principle for the whole realm of physics. He was asked to write a review article 'On the relativity principle and the conclusions drawn from it' [Einstein, 1907]. The reinterpretation of the concept of simultaneity in special relativity was hinging on the finiteness of the speed of light for signal transmission. It was therefore clear that the Newtonian theory of gravitation posed an embarrassment. In Newtonian mechanics, the gravitational force is an action-at-a-distance force and thus contradicts the fundamental assumption of special relativity that no physical effects can propagate with a speed superseding a finite value. In reflection on this difficulty, Einstein took a decisive turn. He linked the problem of the instantaneous propagation of the gravitational force in Newtonian physics to the problem of generalizing the principle of (special) relativity to non-uniform relative motion. In a reinterpretation of Galileo's law of free fall, according to which all bodies in a gravitational field undergo the same acceleration regardless of their weight, Einstein formulated the so-called 'equivalence hypothesis'. According to this hypothesis, there is no conceivable experiment that could distinguish between processes taking place in a static and homogeneous gravitational field and those that are only viewed from a frame of reference that is uniformly and rectilinearly accelerated in a gravitation free space. The value of this hypothesis was a heuristic one. It enabled Einstein to investigate the effects of gravitation in a relativistic theory by analyzing the corresponding processes if interpreted from an accelerated frame of reference.

Already in 1907, Einstein drew three important consequences from the equivalence hypothesis. He concluded that the time and hence also the speed of light must depend on the gravitational potential. Consequently, the frequency of light emitted from the sun should be shifted towards the red, and light rays passing through a gravitational field would be bent. He also concluded that every energy should have not only inertial but also gravitational mass.

Incidentally, this is also the time when Einstein began to use the term 'relativity theory' ('*Relativitätstheorie*') in print; for example, [Einstein, 1907, 439]. The term had first been used in print in the same year by Paul Ehrenfest (1880–1933), after Max Planck (1858–1947) had earlier introduced the term 'Relativtheorie'. A suggestion by Felix Klein (1849–1925) in 1910, to use the perhaps more appropriate term 'invariant theory' ('*Invariantentheorie*'), was not taken up [CPAE, vol. 2, 254].

While the equivalence hypothesis of 1907 provided a point of departure for a generalization of the theory of relativity and for a new field theory of gravitation, Einstein

did not present a solution to the problem of instantaneous propagation of the gravitational force. While he remained rather silent on the topic of the relativity principle for some years, these questions were taken up by others. For example, Hermann Minkowski (1864–1909) and Henri Poincaré (1854–1912) proposed Lorentz-covariant generalizations of Newton's law of gravitation. More importantly, Minkowski also gave the theory of relativity a more sophisticated mathematical representation. Reflecting on the symmetry of the Lorentz transformations in [Minkowski, 1908], he used elements from matrix theory to give the equations a four-dimensional representation and to interpret the Lorentz transformations as rotations in a four-dimensional vector space. In a report of his work to the 80th general assembly of physicians and scientists in Cologne, he illustrated this interpretation by the often-quoted words: 'From this hour on, space by itself, and time by itself, shall be doomed to fade away in the shadows, and only a kind of union of the two shall preserve an independent reality' [Minkowski, 1908, 105]. His four-dimensional representation was taken up by Arnold Sommerfeld (1868–1951) who developed a four-dimensional vector algebra and vector calculus; and by Max Laue (1879–1960) who focused upon the tensorial representation of the stress-energy-momentum complex.

3 THE METRIC TENSOR

Einstein resumed work on the subject again in 1911. By then he had been appointed 'Ordinary' (full) professor of physics at the German University in Prague. In a series of papers he developed a theory of the static gravitational field, following the heuristics of the equivalence assumption of static homogeneous gravitational fields to systems in uniform and rectilinear acceleration [Einstein, 1911, 1912a, 1912b]. His work was boosted by a competition with Max Abraham (1875–1922), who had picked up on Einstein's idea of a variable speed of light and had suggested a dynamic theory of gravitation. Abraham had proposed a field equation where the d'Alembertian acting on the speed of light c was proportional to the scalar mass density. In the course of the debate it quickly became clear that with variable c Abraham's equation was Lorentz covariant at best in some ill-defined infinitesimal sense and could hardly been interpreted consistently. But Abraham had demonstrated to Einstein the technical power of a four-dimensional representation, and had prepared him to take the second big step of introducing the metric tensor.

The second indication of where to go next in the course of generalizing the relativity principle came from the analysis of rotating frames of reference. The heuristic assumption of the equivalence hypothesis implied that also centrifugal and Coriolis forces should be interpreted as gravitational forces. Looking at the invariant

$$dx^2 + dy^2 + dz^2 - c^2\,dt^2 \tag{1}$$

in rotating frames of reference would produce terms of the form $2\omega\,dx\,dt'$, where the angular velocity ω would have to be interpreted as a gravitational potential, just as in the theory of static gravitation the speed of light $c(x, y, z)$ had assumed the role of a variable gravitational potential. Since moreover the measuring rods for determining the circumference, but not the diameter, of a rotating disk are Lorentz contracted, the analysis of a rotating disk already pointed to a breakdown of Euclidean geometry.

4 EINSTEIN'S COLLABORATION WITH MARCEL GROSSMANN

At some point around this time, Einstein remembered Geiser's lectures on Gaussian surface theory that he had studied through the notes of his friend Grossmann. It occurred to him that the invariant line element of differential geometry might be the key to finding a proper mathematical representation for his problem. Fortunately, Einstein had just accepted a call to the Zurich Polytechnic where Grossmann had become professor of geometry in 1907. Einstein asked Grossmann for help in studying the mathematical literature, and the two embarked on an intense collaboration. About this collaboration, he wrote in October 1912 [CPAE, Doc. 421]:

> I am now working exclusively on the gravitation problem and believe that I can overcome all difficulties with the help of a mathematician friend of mine here. But one thing is certain: never before in my life have I troubled myself over anything so much, and I have gained enormous respect for mathematics, whose more subtle parts I considered until now, in my ignorance, as pure luxury.

The question that Einstein put to Grossmann was to identify the mathematics connected with the invariance of a four-dimensional infinitesimal line element

$$ds^2 = \sum_{\mu\nu=1}^{4} g_{\mu\nu}\, dx^\mu\, dx^\nu. \tag{2}$$

A research notebook with calculations from that time documents Einstein's and Grossmann's cooperation [Norton, 1984; Renn and Sauer, 1999; Renn et alii, forthcoming]. It is in this so-called 'Zurich notebook' that we find the first written instance of the metric tensor for $(3+1)$-dimensional space-time ([Renn and Sauer, 1999, 96]; see also Call No. 3-006, image 39, on http://www.alberteinstein.info (2003) for a facsimile). Realizing that the vector calculus for Euclidean space in curvilinear coordinates is formally equivalent to the calculus of a general manifold equipped with an invariant infinitesimal line element, Grossmann saw that the task was to generalize the four-dimensional vector calculus developed by Minkowski, Sommerfeld, Laue, and others using methods of an altogether coordinate independent calculus. Scanning the literature, Grossmann soon found the necessary mathematical concepts in [Riemann, 1892] on n-dimensional manifolds (§39), in [Christoffel, 1869] on quadratic differential forms, and in [Ricci and Levi-Civita, 1901] on their so-called 'absolute differential calculus'.

It seems that Einstein and Grossmann quickly saw how to formulate, in outline, a generally covariant theory with the metric tensor $g_{\mu\nu}$ representing the gravito-inertial field. In the following discussion, I will give all formulas in a notation that is both slightly modernized and made consistent over the various texts discussed. In particular, I will abbreviate coordinate derivatives by subscript commas, use the Einstein summation convention of summing over repeated indices, and denote functional derivatives by δ rather than ∂. In their joint publications, Einstein and Grossmann also used Greek letters to denote contravariant vectors and tensors rather than superscript indices.

Einstein and Grossmann found generally covariant equations of motion of a material point of invariant mass m for a given metric field $g_{\mu\nu}$ in the absence of non-gravitational

forces as

$$\delta\left\{\int L\,dt\right\} = \delta\left\{-m\int ds\right\} = 0, \tag{3}$$

with a particle Lagrangian $L = -m\,ds/dt$. In a generalization to a continuous distribution of matter characterized by an energy-momentum tensor for pressureless flow of dust,

$$T^{\mu\nu} = \rho_0 \frac{dx^\mu}{ds}\frac{dx^\nu}{ds}, \tag{4}$$

the equation of motion turned into

$$\left(\sqrt{-g}g_{\sigma\mu}T^{\mu\nu}\right)_{,\nu} - \frac{1}{2}\sqrt{-g}g_{\mu\nu,\sigma}T^{\mu\nu} = 0, \quad \text{where } g := \det(g_{\mu\nu}). \tag{5}$$

The latter equation is an explicit expression for the vanishing of the covariant divergence of the mixed tensor density $\sqrt{-g}T^\nu_\sigma$. As such it is closely related to the conservation of energy-momentum, as can be seen by integrating $T^{\mu\nu}$ over a closed 3-surface and invoking Gauss's surface theorem. In Einstein's interpretation, the first term of (5) gave the conservation law for special relativity for constant $g_{\mu\nu}$, and the second part consequently represented the energy-momentum flow due to the gravitational field. This interpretation led Einstein to believe that the gravitational force components are given by $g_{\mu\nu,\sigma}$. The task remained to find a field equation for the metric tensor field, i.e. a tensorial generalization of the Poisson equation.

5 COMING CLOSE TO THE SOLUTION, OR SO IT SEEMS

From Riemann's and Christoffel's investigations, Grossmann and Einstein learned that the crucial mathematical concept was the Riemann curvature tensor $\{ik, lm\}$ given in terms of the Christoffel symbols of the second kind (given in the original notation),

$$\left\{\begin{matrix}\mu\ \nu\\ \tau\end{matrix}\right\} = g^{\tau\lambda}(g_{\mu\lambda,\nu} + g_{\nu\lambda,\mu} - g_{\mu\nu,\lambda}), \tag{6}$$

as

$$\{\iota\kappa, \lambda\mu\} = \left\{\begin{matrix}\iota\ \lambda\\ \kappa\end{matrix}\right\}_{,\mu} - \left\{\begin{matrix}\iota\ \mu\\ \kappa\end{matrix}\right\}_{,\lambda} + \left\{\begin{matrix}\iota\ \lambda\\ \rho\end{matrix}\right\}\left\{\begin{matrix}\rho\ \mu\\ \kappa\end{matrix}\right\} - \left\{\begin{matrix}\iota\ \mu\\ \rho\end{matrix}\right\}\left\{\begin{matrix}\rho\ \lambda\\ \kappa\end{matrix}\right\} \tag{7}$$

(see Figure 1). Since the right-hand side of the field equation would be given by the stress-energy tensor of matter, a tensor of second rank, the left-hand side of the field equation also had to be a two-index object. But the obvious candidate, the Ricci tensor

$$R_{\mu\nu} = \{\mu\kappa, \kappa\nu\}, \tag{8}$$

would not produce a field equation that was acceptable to Einstein and Grossmann at the time. Although a field equation,

$$R_{\mu\nu} + \kappa T_{\mu\nu} = 0, \tag{9}$$

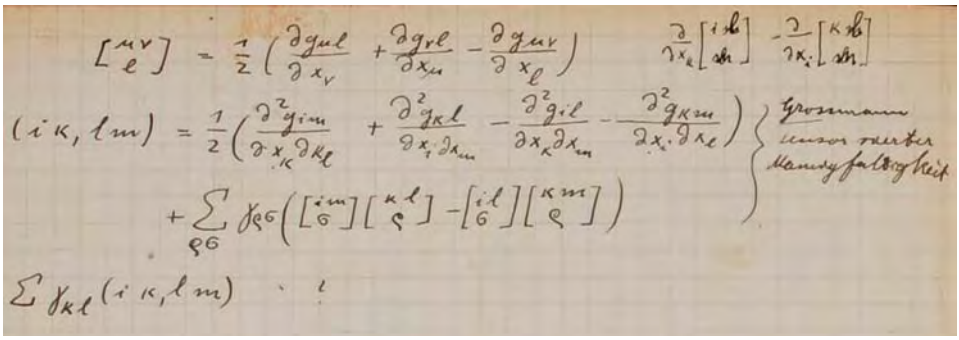


Figure 1. Top portion of page 14L of the 'Zurich Notebook' (Einstein Archives Call No. 3-006). Next to Grossmann's name Einstein writes down the Christoffel symbols of the first kind, and the fully covariant Riemann tensor (ik, lm) which he calls a 'tensor of the fourth manifold' ('Tensor vierter Mannigfaltigkeit'). Einstein then begins to investigate the Ricci tensor by contracting with the contravariant metric γ_{kl}.

with some constant κ was considered as a candidate, they dismissed it because they were unable to recover familiar Newtonian physics in the weak field limit $g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}$ with $\eta_{\mu\nu} = \mathrm{diag}(1, 1, 1, -1)$, $|h_{\mu\nu}| \ll 1$ and $|h_{\mu\nu,\rho}| \ll 1$.

The dismissal of the candidate (9) has been a major puzzle for historians for a long time. Since in the vacuum case $T_{\mu\nu} \equiv 0$, (9) is equivalent to the final field equations of general relativity (see (23) below), Einstein and Grossmann had come by a hair's breadth to arriving at general relativity already at this point, or so it seems. However, a closer analysis of the Zurich notebook revealed that Einstein had to overcome more conceptual difficulties before he was ready to accept a generally covariant theory [Renn and Sauer, 1999; Renn et alii, forthcoming].

6 THE *ENTWURF* THEORY

After giving up the attempt to base a field equation on the Riemann curvature tensor, Einstein and Grossmann constructed a field equation that was closer to their heuristic requirements of energy conservation and recovery of the Poisson equation in the Newtonian limit. The idea was to take the expression $(g^{\alpha\beta} g^{\mu\nu}_{,\beta})_{,\alpha}$, which would clearly reduce to the d'Alembertian and Laplacian operators in the weak field and static limits and substitute it for $T^{\mu\nu}$ in the second term of (5). If additional terms of higher order could be identified such that this expression could be transformed into a total divergence, energy-momentum conservation in the form of (5) would automatically be satisfied. The field equations they found read

$$\frac{1}{\sqrt{-g}}\left(\sqrt{-g}\, g^{\alpha\beta} g^{\mu\nu}_{,\beta}\right)_{,\alpha} - g^{\alpha\beta} g_{\tau\rho} g^{\mu\tau}_{\alpha} g^{\nu\rho}_{\beta} + \frac{1}{2} g^{\alpha\mu} g^{\beta\nu} g_{\tau\rho,\alpha} g^{\tau\rho}_{,\beta} - \frac{1}{4} g^{\mu\nu} g^{\alpha\beta} g_{\tau\rho,\alpha} g^{\tau\rho}_{,\beta} = -\kappa T^{\mu\nu}. \tag{10}$$

In the early summer of 1913, Einstein and Grossmann proceeded to publish their findings in a little booklet under the title *Outline* ['Entwurf'] *of a generalized theory of relativity and of a theory of gravitation* [Einstein and Grossmann, 1913]. As the title page indicated, it was divided into two parts, a physical part for which Einstein signed as responsible, and a mathematical part for which Grossmann signed as author.

The *Entwurf* theory, as it is frequently called in modern historical literature, was a hybrid theory, if viewed from our modern understanding of general relativity. It presented a mathematical apparatus of tensor calculus that allowed to formulate a theory in a generally covariant manner, and it gave generally covariant equations of motion. Just as in the final theory of general relativity, the crucial concept was the metric tensor that was interpreted as representing a gravito-inertial field. All these elements were later to be found in the final version of general relativity. The only thing that was missing were generally covariant field equations.

The hybrid character of the *Entwurf* theory is reflected in a certain ambivalence that Einstein showed with respect to their achievement. Initially, and also again and again over the following two years, he expressed himself rather pleased with the theory. He had settled on the *Entwurf* equations as acceptable equations and began to elaborate their consequences. From an unpublished manuscript we know that together with his friend Michele Besso (1873–1955) he calculated the advance of the planetary perihelia. For Mercury, it was well known that the observed perihelion advance was in discrepancy with the value calculated on the basis of Newtonian mechanics, and this anomaly was the most prominent quantitative failure of classical gravitation theory. Not surprisingly, they found a value for Mercury that was significantly off the observed value: theirs even came with the wrong sign [Earman and Janssen, 1993].

Notwithstanding Einstein's acceptance of the *Entwurf* equations, he also indicated that the restricted covariance of these equations was a 'black spot' of the theory. His initial heuristics clearly did not imply any reason for a restricted covariance of the theory. In further reflection, Einstein convinced himself, however, that this restricted covariance was, in fact, to be expected. He devised an argument to the effect that indeed no generally covariant field equations were physically admissible. The argument was first published in an addendum to a reprint of the *Entwurf* in the *Zeitschrift fur Mathematik und Physik*.

Einstein considered a hole in four-dimensional space-time, i.e. a finite region with vanishing stress-energy $T_{\mu\nu} \equiv 0$. Let $G(x)$ denote a solution $g_{\mu\nu}(x_1, x_2, x_3, x_4)$ of the field equations, and perform a coordinate transformation within the hole, i.e. consider a coordinate system x' that coincides smoothly with the original coordinate system x at the boundary of the hole. In the primed coordinates the transformed field $G'(x')$ is the solution to the transformed field equations. But if the field equations are generally covariant, then $G'(x)$ is also a solution to the original field equations. We hence arrive at two distinct solutions in the same coordinate system x for the same distribution of matter $T_{\mu\nu}$. Einstein concluded that generally covariant field equations cannot uniquely determine the physical processes in a gravitational field. Consequently, one had to restrict the admissible coordinate systems to what he began to call 'adapted coordinates'.

Already in their *Entwurf*, Einstein and Grossmann had stated that the most urgent unsolved problem of their theory was the identification of the covariance group of their field equations. The solution to this question was made possible by a variational reformulation

of the theory. It was the topic of their second joint publication [Einstein and Grossmann, 1913].

As acknowledged in a footnote, the hint of trying a variational approach came from Paul Bernays (1888–1977), a student of David Hilbert (1862–1943) in Göttingen. The idea was that a variational formulation might help to identify the group of 'adapted coordinates' since it would be easier to identify the invariance group of the scalar action integral than the covariance group of the explicit tensorial field equations. Einstein and Grossmann indeed succeeded to cast the *Entwurf* theory in a variational formulation,

$$\delta\left\{\int L\, d^4x\right\} = 0, \tag{11}$$

with a Lagrangian

$$L = \sqrt{-g}\left(\frac{1}{4} g^{\alpha\beta} g_{\tau\rho,\alpha} g^{\tau\rho}_{\beta} - \kappa L^{(\mathrm{mat})}\right), \tag{12}$$

where the matter part $L^{(\mathrm{mat})}$ was not included explicitly.

Considering variations adapted to the hole consideration, they were now able to identify the condition for 'adapted coordinates' governing the covariance group of the *Entwurf* as

$$B_\sigma = \left(\sqrt{-g} g^{\alpha\beta} g_{\sigma\mu} g^{\mu\nu}_{,\beta}\right)_{,\nu\alpha} = 0. \tag{13}$$

With their second joint *paper*, the collaboration between Einstein and Grossmann came to an end. In spring 1914, Einstein moved to Berlin taking up a position as a member of the Prussian Academy of Sciences in Berlin, a move that relieved him of his teaching load as professor at the Zurich Polytechnic.

7 THE 1914 REVIEW ARTICLE ON THE *ENTWURF* THEORY

In summer 1914, Einstein felt that the new theory should be presented in a comprehensive review. He also felt that a mathematical derivation of the field equations that would determine them uniquely was still missing.

Both tasks are addressed in a long paper, presented in October 1914 to the Prussian Academy for publication in its *Sitzungsberichte* [Einstein, 1914]. It is entitled 'The formal foundation of the general theory of relativity'; here, for the first time, Einstein gave the new theory of relativity the epithet 'general' in lieu of the more cautious 'generalized' that he had used for the *Entwurf*.

The paper is divided into five sections, and thus anticipates the structure of the final 1916 review. An introductory section on the basic ideas of the theory is followed by a section on the theory of covariants. This section replaced Grossmann's mathematical part of the joint *Entwurf* paper and gives an account of the elements of tensor calculus employed in the theory. A third section discusses the theory for a given metric field. It introduced the stress-energy-momentum tensor and discussed the conservation laws associated with the vanishing of its divergence, as well as the equations of motions and the electromagnetic field equations.

The fourth section gave a new derivation of the *Entwurf* equations. Einstein here tried to give a derivation that supposedly rendered them unique. He reiterated the hole consideration and introduced adapted coordinates. The variation is now done in a generic manner for the gravitational part H of the Lagrangian L. In order to fix the Lagrangian, Einstein assumes to be a homogeneous function of second degree in the coordinate derivatives $g^{\mu\nu}_{,\sigma}$ of the metric, and picks from the allowed combinations the one that conforms to the adapted coordinate condition.

In a final, short section Einstein discussed approximations of the theory, recovered the Newtonian limit and predicted both gravitational light bending and red shift.

8 THE DEMISE OF THE *ENTWURF* AND THE BREAKTHROUGH TO GENERAL COVARIANCE

Einstein had known that the *Entwurf* equations produced the wrong perihelion advance for Mercury since 1913. A second set-back that undermined his confidence in the theory came in spring 1915 when Tullio Levi-Civita (1873–1941) carefully studied Einstein's long Academy paper and found fault with its derivation of the field equations. After an intense epistolary exchange in March and April 1915, Einstein had to admit that his proof of the tensorial character of the left hand side of the field equations for admissible coordinate transformations was incomplete [CPAE, vol. 8, Doc. 80].

In September 1915, Einstein realized that the Minkowski metric in rotating Cartesian coordinates is not a solution to the *Entwurf* equations. Earlier checks of this condition appear to have been flawed by trivial algebraic mistakes that conspired to convince him of the validity of this heuristic requirement [Janssen, 1999]. The final blow came quickly afterwards when Einstein discovered that the alleged uniqueness of the field equations in his derivation of the Academy paper did not hold up.

At this point, Einstein began to reconsider alternatives for the gravitational field equations. He reflected on considerations that he had done previously in his search for the *Entwurf* equations. A closer analysis of the Zurich notebook indeed revealed that in the fall of 1915, Einstein reconsidered the same candidates for field equations as he had done in 1912 [Norton, 1984; Renn and Sauer, 1999; Renn et alii, forthcoming]. The return to general covariance is documented in four communications to the Prussian Academy, presented on 4, 11, 18 and 25 November, and each published a week later in the *Sitzungsberichte*.

In the first communication, Einstein announced that he had lost his faith in the *Entwurf* equations and wrote: 'In this pursuit I arrived at the demand of general covariance, a demand from which I parted, though with a heavy heart, three years ago when I worked together with my friend Grossmann. As a matter of fact, we were then quite close to that solution of the problem, which will be given in the following' [Einstein, 1915a, 778].

Einstein now split the Ricci tensor into two parts,

$$R_{\mu\nu} = \{\mu\kappa, \kappa\nu\} = N_{\mu\nu} + M_{\mu\nu}, \tag{14}$$

where

$$N_{\mu\nu} = -\left\{ \begin{matrix} \mu & \nu \\ \kappa & \end{matrix} \right\}_{,\kappa} + \left\{ \begin{matrix} \mu & \kappa \\ \rho & \end{matrix} \right\} \left\{ \begin{matrix} \rho & \nu \\ \kappa & \end{matrix} \right\}, \tag{15}$$

and

$$M_{\mu\nu} = -\left\{ \begin{matrix} \mu\ \kappa \\ \kappa \end{matrix} \right\}_{,\nu} + \left\{ \begin{matrix} \mu\ \nu \\ \rho \end{matrix} \right\} \left\{ \begin{matrix} \rho\ \kappa \\ \kappa \end{matrix} \right\}. \tag{16}$$

Since $\left\{ \begin{matrix} \mu\ \kappa \\ \kappa \end{matrix} \right\} = (\ln \sqrt{-g})_{,\mu}$ is a vector for all transformations that leave g invariant (unimodular substitutions), $M_{\mu\nu}$ is a covariant derivative of a vector, and hence all quantities in (14) are tensors under such substitutions.

The field equations of the first November communication were now given as

$$N_{\mu\nu} = -\kappa T_{\mu\nu}. \tag{17}$$

Even though Einstein explicitly reverted to the general covariance of the Riemann–Christoffel tensor, the field equations of this communication are not generally covariant, but only under unimodular coordinate transformations.

The restricted covariance is immediately obvious also from the variational formulation that Einstein provided. Looking again at the geodesic equation,

$$\frac{d^2x^\tau}{ds^2} + \left\{ \begin{matrix} \mu\ \nu \\ \tau \end{matrix} \right\} \frac{dx^\mu}{ds} \frac{dx^\nu}{ds} = 0, \tag{18}$$

as the equation of motion for a point particle in a given gravitational field, he now conceived of the negative Christoffel symbols $\Gamma^\sigma_{\mu\nu} = -\left\{ \begin{matrix} \mu\ \nu \\ \sigma \end{matrix} \right\}$ as the components of the gravitational force rather than the simple coordinate derivatives of the metric $g_{\mu\nu,\sigma}$. These quantities now entered into the gravitational part of the Lagrangian as

$$L = g^{\sigma\tau} \Gamma^\alpha_{\sigma\beta} \Gamma^\beta_{\tau\alpha} - \kappa L^{(\mathrm{mat})}. \tag{19}$$

(compare (12)). He observed that weak fields now allow to go to the Newtonian limit, and that the transition to rotating frames of reference is admissible since the corresponding coordinate transformations have unit determinant.

Not only was the covariance of the theory restricted to unimodular transformations; Einstein also showed that energy-momentum conservation demanded that a coordinate restriction,

$$\left(g^{\alpha\beta} \left[\ln \sqrt{-g} \right]_{,\beta} \right)_{,\alpha} = -\kappa T, \tag{20}$$

had to be satisfied. Since, in general, the trace of the energy-momentum tensor $T = g^{\mu\nu} T_{\mu\nu}$ does not vanish, (20) implies that coordinates cannot be chosen arbitrarily. In particular, (20) implies that one cannot set $\sqrt{-g} \equiv 1$.

At this point, it needs to be mentioned that Einstein's return to general covariance in November 1915 was done in a hasty competition with Hilbert [Sauer, 1999]. Einstein had given a series of lectures on the *Entwurf* theory in Gottingen earlier in the summer, and Hilbert had then closely studied Einstein's theory over the fall. Apparently, Hilbert had found fault with Einstein's derivation of the field equations, too, and Einstein had heard about Hilbert's criticism through Sommerfeld [CPAE, vol. 8, Doc. 136]. When he received

proofs of his first November communication, he forwarded them to Göttingen, and it seems that Hilbert responded immediately with a report about his own progress. At the time Hilbert believed in an electromagnetic world-view and had been working on combining Einstein's gravitational theory with a generalized version of Maxwellian electrodynamics suggested by Gustav Mie (1868–1957). Mie had proposed a theory of matter where non-linear, but Lorentz-covariant generalizations of Maxwell's equations should allow for particle-like solutions in the microscopic realm. It seems likely that Hilbert had informed Einstein about the basic characteristics of his approach which aimed at a unification of Einstein's and Mie's theories.

The second of Einstein's four famous November communications, in any case, discussed the possibility of a purely electromagnetic origin of matter [Einstein, 1915b]. Since in classical electromagnetism, the stress-energy-momentum tensor $T^{\mu\nu}$ is given in terms of the electromagnetic field tensor $F_{\mu\nu}$ as

$$T^{\mu\nu} = \frac{1}{4\pi}\left(F^{\mu\alpha}F^{\nu}_{\alpha} - \frac{1}{4}g^{\mu\nu}F^{\alpha\beta}F_{\alpha\beta}\right), \tag{21}$$

it is readily seen that its trace T vanishes identically. Einstein now entertained the possibility that on a microscopic level all matter might be of electromagnetic origin. In this case, the right-hand side of the coordinate condition (20) would vanish and hence coordinates with constant g would be admissible. In this case, Einstein argued, one could take the fully covariant equations

$$R_{\mu\nu} = -\kappa T_{\mu\nu}, \tag{22}$$

that he had already considered earlier in (9) and reduce them to the field equations (17) by choosing coordinates for which $g \equiv 1$.

The field equations (22) still differ from the final field equations, but for the vacuum case, $T_{\mu\nu} = 0$, they are already equivalent. Einstein therefore was able to compute on the basis of (22) the correct unaccounted perihelion advance by looking at the field of a point mass in second approximation. The calculation produced the correct value of $43''$ per century without any arbitrary or ad hoc assumptions. In the computation Einstein could take advantage of his having calculated the advance before for the *Entwurf* theory. The new field equations, in fact, only involved a modification of his earlier calculations [Earman and Janssen, 1993]. Einstein published these results in his third November communication [Einstein, 1915c].

With the success of the perihelion calculation, the return to general covariance was definite. The final step [Einstein, 1915d] was to add a trace term to the matter tensor to obtain field equations of the form

$$R_{\mu\nu} = -\kappa\left(T_{\mu\nu} - \frac{1}{2}g_{\mu\nu}T\right). \tag{23}$$

With the trace term added, the postulate of energy-momentum conservation no longer produced a coordinate restriction since it was now automatically satisfied by (23).

Equations (23) are the final field equations of the generally relativistic theory of gravitation, as we know them today. They are frequently referred to as the 'Einstein equations' of general relativity.

With the exception of the first November communication, where he had given the Lagrangian (19) for the field equations (17), Einstein had not discussed the subsequent field equations in a variational approach. The closure of providing a variational formulation was contributed by Hilbert in his own approach to a generally covariant theory of gravitation and electromagnetism [Hilbert, 1915]. Since he was being kept informed by Einstein about the latter's progress, he rushed ahead and presented an account of his own version to the Göttingen Academy for publication in its *Nachrichten* on November 20. Page proofs of Hilbert's original paper show that the version submitted for publication on November 20 still differed from the version that was eventually published. But it did already suggest to base the theory on a variational principle and emphasized that the Lagrangian must be a scalar function for general coordinate transformations.

In the printed version of Hilbert's paper, the Riemann curvature scalar R is taken to be the gravitational part of the Lagrangian; and it is stated, albeit not derived by explicit calculation, that a variation of the action

$$\mathcal{A} = \int \sqrt{-g}\big(R - \kappa L^{(\mathrm{mat})}\big)\, d^4\tau, \tag{24}$$

with respect to the metric tensor components $g^{\mu\nu}$ would produce the gravitational field equations

$$R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R = -\kappa \frac{1}{\sqrt{-g}} \frac{\delta L^{(\mathrm{mat})}}{\delta g^{\mu\nu}}, \tag{25}$$

which is an equivalent version of Einstein's field equation (23). (25) may be transformed to (23) by looking at the trace of (23) and substituting $R = -\kappa T$ into it. The equivalence then follows from the non-trivial identification of

$$\frac{1}{\sqrt{-g}} \frac{\delta L^{(\mathrm{mat})}}{\delta g^{\mu\nu}} = T_{\mu\nu}. \tag{26}$$

In the latter step, Hilbert and Einstein differed considerably since Hilbert axiomatically took $L^{(\mathrm{mat})}$ to be a function exclusively of the electromagnetic potential A_μ, the electromagnetic field $F_{\mu\nu} = A_{\mu,\nu} - A_{\nu,\mu}$, and the metric tensor components $g_{\mu\nu}$,

$$L^{(\mathrm{mat})} = L^{(\mathrm{mat})}(A_\mu, F_{\mu\nu}, g_{\mu\nu}), \tag{27}$$

in accordance with his electromagnetic world view. Einstein, however, had entertained the hypothesis of an electromagnetic origin of matter only for a few days. With his fourth November communication at the latest, he had given up that hypothesis again and was allowing for an unspecified $T_{\mu\nu}$ in his final version of the theory.

9 THE 1916 REVIEW PAPER

Ever since Levi-Civita had found a gap in Einstein's covariance proof of the *Entwurf* equations, Einstein had meant to update or rewrite his 1914 Academy article on the general theory of relativity. With the return to general covariance, the success of explaining the perihelion advance of Mercury, and the new field equations (23) of the fourth November communication, he decided to write an altogether new account of the general theory of relativity.

The new review was received by the *Annalen der Physik* on 20 March 1916, some four months after the last November paper. It is the landmark subject of this article. Its structure is not much different from the earlier 1914 Academy article. It is again divided into five Sections:

[A.] Fundamental considerations on the postulate of relativity;

[B.] Mathematical aids to the formulation of generally covariant equations;

[C.] Theory of the gravitational field;

[D.] Material phenomena;

[E.] [Newtonian limit and observable consequences].

In an introductory paragraph Einstein called the theory to be expounded in the review 'conceivably the farthest-reaching generalization' of the special theory of relativity. While the latter is assumed to be known to the reader, he sets out to develop especially all the necessary mathematical tools'—and I tried to do it in as simple and transparent a manner as possible, so that a special study of the mathematical literature is not required for the understanding of the present paper' (p. 769).

Nevertheless, in this first paragraph Einstein did mention Minkowski's formal equivalence of the spatial and time coordinates, the investigations on non-Euclidean manifolds by Gauss, Riemann, and Christoffel, and the absolute differential calculus of Ricci and Levi-Civita. Echoing a theme of Felix Klein's but also of later commentators, he wrote that especially the absolute differential calculus had provided mathematical means which simply had to be taken up—as if he had not struggled hard for years to apply them in a physically meaningful way. He also acknowledged Grossmann's help again in studying the mathematical literature and in searching for the gravitational field equations.

The first Section then introduces the postulate of general covariance, arguing to a large extent from purely epistemological considerations. Einstein denounces the existence of an absolute space by considering two massive bodies far away both from other masses and from each other and in relative rotation along their line of connection. If one body were observed to be of spherical shape and the other to be an ellipsoid, then Newtonian mechanics would have to attribute the cause for the different shapes in a rotation relative to absolute space. But this is unsatisfactory because a causal agent is introduced which itself can never be an object of causal effect nor of observation. Hence, one is forced to attribute the cause for this change of shape to the distant masses of the fixed stars, an argument that follows Mach's critique of classical mechanics.

The second argument is the equivalence hypothesis based on Galileo's empirical law of free fall. Next, Einstein discusses the rotating disk to argue for the fact that in general relativity coordinates no longer have an immediate metric meaning. A fourth argument in this Section was new and replaced the earlier hole consideration. The hole argument had supposedly proven that no generally covariant field equations could be given a physical meaning in accordance with our notions of causality and the demand that the field equations are determined uniquely by the energy-matter distribution. Einstein did not explicitly retract the argument but gave a new consideration, known as the point coincidence argument. He argued that what we observe in physical experiments are always only spatio-temporal coincidences. If all physical processes would consist in the motion of material points, we could only observe those events where two or more of their worldlines coincide. Then the coordinates of the four-dimensional space-time manifold are merely labels for those coincidences, and no coordinate system must be preferred over any other. The implicit objection to the hole argument that invalidates its conclusion is that the different metric fields $G(x)$ and $G'(x)$ obtained by dragging the metric tensor over the hole, do not, in fact, represent different physical situations since they agree on all point coincidences.

In the second, mathematical Section, Einstein summarily develops the elements of tensor algebra and tensor calculus. He introduces contravariant and covariant vectors and general tensors that are defined by the transformation laws of their components. He introduces the algebraic operations of external multiplication and contraction, and of raising and lowering of indices. Among the properties of the metric tensor, he discusses the invariance of the volume element $\sqrt{-g}\,d^4x$. He repeats the derivation of the geodesic equation, introduces Christoffel's symbols and discusses covariant differentiation by considering invariance along the geodesic line. He mentions the fact that the covariant derivative of the metric vanishes and derives a number of explicit formulas for the differentiation of contravariant, covariant and mixed tensors. The last paragraph introduces the Riemann–Christoffel curvature tensor and discusses its splitting into two parts, as in (14). Perhaps the most noteworthy point of the Section, compared to earlier expositions of the mathematical foundations of general relativity, is what came to be called the 'Einstein summation convention'. It is in this section that for the first time in print he introduced the convention that in any tensor expression a summation over two repeated indices is implied without writing down the summation sign.

The third Section derives the gravitational field equations. They are given as

$$\Gamma^{\alpha}_{\mu\nu,\alpha} + \Gamma^{\alpha}_{\mu\beta}\Gamma^{\beta}_{\nu\alpha} = -\kappa\left(T_{\mu\nu} - \frac{1}{2}g_{\mu\nu}T\right), \tag{28}$$

$$\sqrt{-g} = 1. \tag{29}$$

Somewhat surprisingly, from a modern point of view, Einstein did not give the field equations in a generally covariant form. Instead, he fixed the coordinates by condition (29) in all equations that he gave in the Section. He emphasized, though, that this is a mere specification of the coordinates introduced for convenience. The introduction of the field equations, in fact, proceeded by arguing that the vanishing of the Ricci tensor $R_{\mu\nu}$ is the unique equation that determines the metric field in the absence of masses if we demand that the expression depends only on $g_{\mu\nu}$ and its first and second derivatives and moreover

on the latter terms only linearly. The possibility of adding a term proportional to $g_{im}R$, equivalent in the vacuum case (but not of adding a cosmological term proportional to g_{im}), is mentioned in a footnote.

The Lagrangian for the variational form of the field equations in vacuum is given as

$$L = g^{\mu\nu}\Gamma^{\alpha}_{\mu\beta}\Gamma^{\beta}_{\nu\alpha}, \tag{30}$$

together with the explicit stipulation of condition (29). The introduction of the matter term proceeds by defining the stress-energy complex of the gravitational field as

$$\kappa t^{\alpha}_{\sigma} = \frac{1}{2}\delta^{\alpha}_{\sigma} g^{\mu\nu}\Gamma^{\alpha}_{\mu\beta}\Gamma^{\beta}_{\nu\alpha} - g^{\mu\nu}\Gamma^{\alpha}_{\mu\beta}\Gamma^{\beta}_{\nu\sigma}, \tag{31}$$

an expression that is not a tensor under general coordinate transformation, in accordance with the fact that the field energy associated with the gravito-inertial field is not a localizable quantity. Using t^{α}_{σ}, Einstein rewrote the field equation (28) as

$$\left(g^{\sigma\beta}\Gamma^{\alpha}_{\mu\beta}\right)_{,\alpha} = \kappa\left(t^{\sigma}_{\mu} - \frac{1}{2}\delta^{\sigma}_{\mu}t\right), \tag{32}$$

and demanded that the non-gravitational energy-momentum tensor T^{σ}_{μ} enters in the equation on the same footing as t^{σ}_{μ}. The latter requirement is equivalent to demanding that a divergence equation,

$$\left(t^{\sigma}_{\mu} + T^{\sigma}_{\mu}\right)_{,\sigma} = 0, \tag{33}$$

holds for the total energy of the system.

While the derivation of the field equations differs considerably from earlier accounts, the fourth and fifth Sections take up material from earlier expositions. In these sections, Einstein discussed Euler's hydrodynamic equation with an energy-momentum tensor

$$T^{\alpha\beta} = -g^{\alpha\beta}p + \rho u^{\alpha}u^{\beta}, \quad \text{where } u^{\alpha} = dx^{\alpha}/ds, \tag{34}$$

for non-dissipative, adiabatic liquids, characterized by the two scalars of pressure p and density ρ. Electrodynamics is governed by Maxwell's equations in generally covariant form, and, in the last Section, Einstein discussed the Newtonian approximation of weak fields, Minkowski flat boundary conditions, and slow motion of the particles. In the consideration of the Newtonian limit the constant may be related to the gravitational constant G by comparison with Poisson's equation as $\kappa = 8\pi G/c^2$. Einstein explains a subtlety of the Newtonian limit that had played a role in his earlier dismissal of generally covariant equations. In the first Newtonian approximation only the g_{44} components enter into the equations of motion, even though the postulate $\sqrt{-g} = 1$ demands that the other diagonal components are non-trivial of the same order. The first-order diagonal components, however, do enter into the geodesic equation for a light ray passing in a centrally symmetric gravitational field. For this reason, the predicted expression for the light bending of a light ray grazing the edge of the Sun, came out with a factor of 2, compared to earlier considerations that were based on the equivalence hypothesis alone. The slowing of clocks in a gravitational field and the gravitational red shift of spectral lines is discussed explicitly, but

the calculation of the perihelion shift for Mercury obtained in second approximation of a spherically symmetric field is only mentioned with reference to the pertinent November communication.

10 EARLY RECEPTION OF THE FINAL VERSION OF GENERAL RELATIVITY

The first exact solution to the field equations—and to date the most important one—was found almost immediately after Einstein had published his field equations (23) by the astronomer Karl Schwarzschild (1873–1916). He computed the field first of a mass point and then of a spherically symmetric mass distribution of total mass m. His solution allowed to compute the light bending of light rays and the planetary perihelion motion without approximation [Schwarzschild, 1916]. The solution is regular everywhere except at the origin but at a radius $r_S = 2Gm/c^2$, now called the Schwarzschild radius, the time coordinate changes its sign relative to the spatial coordinates. This coordinate singularity is responsible for what came to be known as the black hole horizon and its interpretation presented a major difficulty for many years.

While more exact solutions were found over the following years, approximation schemes played an equally important role for an interpretation of the theory. An approximate solution was discussed by Einstein in the summer of 1916 in a first paper on gravitational waves. The existence of gravitational waves was expected in a field theory of gravitation by analogy to the electromagnetic case. Einstein's first paper on this topic was marred by a mistake which made him conclude that waves should exist that do not transport energy. The error was corrected in a second paper of 1918. Until now, the topic of gravitational waves is an active field of research and their existence has been shown indirectly only in 1974 through the energy loss of binary pulsars (Nobel prize 1993). Experimental efforts to observe gravitational waves directly are still underway.

The question of energy transport in gravitational waves is connected to the question of identifying an expression for the gravitational field energy and a corresponding conservation law. The question was debated in the years 1916–1919 by a number of mathematicians, most importantly by Felix Klein. The final solution came with Noether's theorems on the connection of conservation laws and symmetries of the variational formulation. These theorems were anticipated for a special case in Hilbert's 1915 paper and published in its general form in 1918 by Emmy Noether (1882–1935).

Einstein tried to encourage experimental efforts aimed at testing the two main predictions of the theory. A confirmation of the gravitational red shift was difficult to determine due to the many competing effects that result in a shifting or broadening of solar or stellar spectral lines. An unequivocal confirmation of the gravitational red shift only came in 1960 in a controlled terrestrial experiment making use of the Mossbauer effect (concerning the gamma-ray spectrum).

But the results of a British expedition led by Arthur Eddington (1882–1944) to test the predicted gravitational light bending during a solar eclipse on 29 May 1919 in Sobral, Brazil, and on the island of Principe in the Gulf of Guinea, reached Europe later in the fall of that year. The results confirmed Einstein's prediction, and within weeks of their publication in the popular media, Einstein turned into a world celebrity and the theory of relativity into a household term.

A popular, non-technical account of both the special and general theories of relativity that Einstein had written as [Einstein, 1917] became a best-seller. A fourth edition in 1919 was reprinted in a fifth through tenth edition in 1920 and saw a fourteenth edition in 1922. It was also translated into many languages. The increased interest in Einstein's theory is also witnessed by an uncountable number of more or less popular accounts and other books and articles dealing with relativity. A bibliography of relativity from 1924 lists close to 4000 entries [Lecat, 1924].

The consequences of both special and general relativity began to be discussed in many circles. Early interpretations of general relativity from a philosophical point of view had been published by Moritz Schlick (1882–1936) and Hans Reichenbach (1891–1953). In the early 1920s philosophical interpretations of relativity came to abound; the analysis in [Hentschel, 1990] carries a bibliography of over 3000 items. The public interest in Einstein's new theory was not always untainted by political partisanry. Antisemitic attacks against Einstein not only focussed on his person or on his political and pacifist stance but also targeted his theory as well. As early as 1920, antisemitically motivated objections against the theories of relativity were expressed in a public meeting at the Berlin Philharmonic in summer 1920, and again at the first post-war meeting of the Society of German Scientists and Physicians in Bad Nauheim in September 1920. On the other hand, Einstein began to be recognized worldwide as a leading physicist. He received international invitations and honors, and began to travel extensively giving talks about his theory at a time when post-war German science was still boycotted by many scholars and scientific institutions.

11 GOING ON AND BEYOND GENERAL RELATIVITY

For Einstein, the victory of the breakthrough to general covariance in November 1915 was not to be regarded as establishing a final theory that would not be subject to further revisions. Already in 1917, he modified the gravitational field equations by adding a term proportional to $\lambda g_{\mu\nu}$ to (23). The modification was motivated in the context of a cosmological consideration. Einstein wanted to avoid the stipulation of boundary conditions at infinity in order not to have to account for inertial effects that might not have been caused by masses, in accordance with what he called Mach's principle. He suggested to consider the cosmological model of a spatially closed and static universe but had to modify the field equations by introducing the cosmological constant λ in order to allow for the possibility of such a solution. An alternate vacuum solution to the modified field equations advanced by Willem de Sitter (1872–1934) soon showed, however, that the new field equations did not automatically satisfy Mach's principle as had been Einstein's hope.

In 1919, Einstein entertained the possibility of a gravitational field equation where the trace term in (23) would be added with a factor of $1/4$ instead of $1/2$. The modification was motivated by considerations concerning the constitution of matter and implies that it is no longer the covariant divergence of $T_{\mu\nu}$ that is automatically vanishing but rather its trace. Other modifications of the field equations or generalizations of the underlying Riemannian geometry were investigated by Einstein and others in the following decades in attempts to find a geometrized unification of the gravitational and electromagnetic fields.

In fact, a geometric interpretation of the general theory of relativity, if considered at all, originally pertained only to the geodesic equation. Until 1916, the Riemann and Ricci tensors were only interpreted as algebraic invariants. A geometric interpretation in terms of parallel transport of tangent vectors was elaborated in the following years mainly through the work of Levi-Civita and Hermann Weyl (1885–1955).

In the course of elaborating the geometric meaning of general relativity, it was Hermann Weyl who took the first steps to go beyond a purely (semi-)Riemannian framework for general relativity and, at the same time, first proposed a truly geometrized unification of the gravitational and electromagnetic fields. First published independently, it was also incorporated into the third edition of his widely read exposition of general relativity ([Weyl, 1918]; see [Scholz, 2001]). In accordance with more general philosophical concerns about the foundations of mathematics, Weyl's point of departure was the observation that in Riemannian geometry, no integrable, or path-independent comparison of vector directions at different points of the manifold is possible, whereas the length of a vector remains unaffected during parallel transport. In order to realize a true 'infinitesimal geometry' (*Nahegeometrie*), Weyl introduced an additional geometric structure, a length connection, i.e. a linear differential form φ that governed the transport of vector lengths l by the definition

$$\delta l \equiv (\partial l/\partial x^i)\, dx^i + l\phi_i\, dx^i \equiv 0. \tag{35}$$

At the same time, the Riemannian metric $g_{\mu\nu}$ had to be replaced by the class of conformally equivalent metrics $[g]$, where two representatives of a class are connected through $\tilde{g}_{\mu\nu} = \lambda g_{\mu\nu}$ with a scalar function λ. For consistency, the length connection φ has to be transformed, too, as $\tilde{\varphi}_i\, dx^i = \varphi_i\, dx^i - d\log\lambda$. For these transformations, Weyl introduced the term 'gauge transformations'.

The (semi-)Riemannian manifold with metric tensor field $g_{\mu\nu}$ was hence generalized to a manifold with conformally equivalent classes $[g]$, $[\varphi]$ of (semi-)Riemannian metrics and length connections. The geometric meaning of this generalization was realized by investigating the affine connection, governing the parallel transport of vectors. It turned out that the curvature associated with the length connection, i.e. the exterior derivative of $f = d\varphi$ (in coordinates, $f_{ij} = \varphi_{i,j} - \varphi_{j,i}$) could be interpreted as the representation of the electromagnetic field tensor [Scholz, 2001, esp. pp. 63–69].

Einstein's reaction to Weyl's theory was highly ambivalent. Fascinated by the mathematical analysis, he quickly pointed out that the theory was unacceptable from a physics point of view since it implied, for example, that the wavelength of light emitted by radiating atoms should depend on the prehistory of the atom, contrary to experience. Despite this argument, Weyl's theory proved extremely influential as the first (more or less) successful attempt to achieve a geometric unification of the gravitational and electromagnetic fields. During the 1920s, many attempts were tried to achieve a unification of gravitation and electromagnetism by generalizing Riemann geometry. These investigations both stimulated and profited from parallel developments in differential geometry.

With the advent of quantum mechanics in 1926, the discovery of the weak and strong interactions and the proliferation of elementary particles in nuclear and subnuclear physics, the parameters for a unification program changed drastically (compare §69). Many aspects of the original unified field theory program have consequently fallen into oblivion, but

the history of modern differential geometry can hardly be understood without taking into account this context of searching for generalizations of Riemannian geometry.

In essence, Einstein's general theory of relativity of 1916 remains today the accepted theory of the gravitational field, and notwithstanding the expectation that a generally relativistic theory of gravitation should also be quantized—still an unsolved problem—classical general relativity, in the sense of an exploration of the solutions and implicit consequences of its gravitational field equations, has been an active field of research ever since.

BIBLIOGRAPHY

CPAE. *The collected papers of Albert Einstein*, Princeton: Princeton University Press, 1987ff.

Christoffel, E.B. 1869. 'Ueber die Transformation der homogenen Differentialausdrücke zweiten Grades', *Journal für die reine und angewandte Mathematik, 70*, 46–70.

Earman, J. and Janssen, M. 1993. 'Einstein's explanation of the motion of Mercury's perihelion', in [Earman, Janssen and Norton, 1993], 129–172.

Earman, J., Janssen, M. and Norton, J. (eds.) 1993. *The attraction of gravitation*: *new studies in the history of general relativity*, Boston, Basel and Berlin: Birkhäuser (Einstein studies, vol. 5).

Einstein, A. 1905. 'Zur Elektrodynamik bewegter Körper', *Annalen der Physik, 17*, 891–921. [Repr. in CPAE, vol. 2, 275–310.]

Einstein, A. 1907. 'Über das Relativitätsprinzip und die aus demselben gezogenen Folgerungen', *Jahrbuch der Radioaktivität und Elektronik, 4*, 411–462. [Repr. in CPAE, vol. 2, 432–488.]

Einstein, A. 1911. 'Über den Einfluss der Schwerkraft auf die Ausbreitung des Lichtes', *Annalen der Physik, 35*, 898–908. [Repr. in CPAE, vol. 3, 485–497.]

Einstein, A. 1912a. 'Lichtgeschwindigkeit und Statik des Gravitationsfeldes', *Ibidem, 38*, 355–369. [Repr. in CPAE, vol. 4, 129–145.]

Einstein, A. 1912b. 'Zur Theorie des statischen Gravitationsfeldes', *Ibidem, 38*, 443–458. [Repr. in CPAE, vol. 4, 146–164.]

Einstein, A. 1914. 'Die formale Grundlage der allgemeinen Relativitätstheorie', *Sitzungsberichte der Königlichen Preussischen Akademie der Wissenschaften* (*Berlin*), 1030–1085. [Repr. in CPAE, vol. 6, 72–130.]

Einstein, A. 1915a, 1915b. 'Zur allgemeinen Relativitätstheorie' and '(Nachtrag)', *Ibidem*, 778–786, 799–801. [Repr. in CPAE, vol. 6, 214–224, 225–229.]

Einstein, A. 1915c. 'Erklärung der Perihelbewegung des Merkur aus der allgemeinen Relativitätstheorie', *Ibidem*, 831–839. [Repr. in CPAE, vol. 6, 233–243.]

Einstein, A. 1915d. 'Die Feldgleichungen der Gravitation', *Ibidem*, 844–847. [Repr. in CPAE, vol. 6, 244–249.]

Einstein, A. 1917. *Über die spezielle und die allgemeine Relativitätstheorie*, Braunschweig: Vieweg. [Repr. in CPAE, vol. 6, 420–539.]

Einstein, A. and Grossmann, M. 1913. *Entwurf einer verallgemeinerten Relativitätstheorie und einer Theorie der Gravitation*, Leipzig: Teubner. [Repr. with addendum in *Zeitschrift für Mathematik und Physik, 62*, 225–261; also in CPAE, vol. 4, 302–343, 579–582.].

Einstein, A. 1914 'Kovarianzeigenschaften der Feldgleichungen der auf die verallgemeinerte Relativitätstheorie gegründeten Gravitationstheorie', *Zeitschrift fur Mathematik und Physik, 63*, 215–225. [Repr. in CPAE, vol. 6, 6–18.]

Eisenstaedt, J. and Kox, A.J. (eds.) 1992. *Studies in the history of general relativity*, Boston, Basel and Berlin: Birkhäuser (Einstein Studies, vol. 3).

Fölsing, A. 1998. *Albert Einstein. A biography*. Harmondsworth: Penguin.

Goenner, H., Renn, J., Ritter, J. and Sauer, T. (eds.) 1999. *The expanding worlds of general relativity*, Boston, Basel and Berlin: Birkhäuser (Einstein Studies, vol. 7).

Hentschel, K. 1990. *Interpretationen und Fehlinterpretationen der speziellen und der allgemeinen Relativitätstheorie durch Zeitgenossen Albert Einsteins*, Basel, Boston and Berlin: Birkhäuser.

Hilbert, D. 1915. 'Die Grundlagen der Physik. (Erste Mitteilung.)', *Nachrichten der Königlichen Gesellschaft der Wissenschaften zu Göttingen, mathematisch-physikalische Klasse*, 395–407.

Howard, D. and Stachel, J. (eds.) 1989. *Einstein and the history of general relativity*, Boston, Basel and Berlin: Birkhäuser (Einstein Studies, vol. 1).

Janssen, M. 1999. 'Rotation as the nemesis of Einstein's *Entwurf* theory', in [Goenner et alii, 1999], 127–157.

Lecat, M. 1924. *Bibliographie de la relativité*, Brussels: Lamertin.

Minkowski, H. 1908. 'Die Grundgleichungen fur die elektromagnetischen Vorgänge in bewegten Körpern', *Nachrichten der Königlichen Gesellschaft der Wissenschaften zu Göttingen, mathematisch-physikalische Klasse*, 53–111.

Minkowski, H. 1909. 'Raum und Zeit', *Physikalische Zeitschrift, 10*, 104–111.

Norton, J. 1984. 'How Einstein found his field equations', *Historical studies in the physical sciences, 14*, 253–316. [Repr. in [Howard and Stachel, 1989], 101–159.]

Pais, A. 1982. *'Subtle is the Lord…'. The science and the life of Albert Einstein*, Oxford: Oxford University Press.

Ricci, G., and Levi-Civita, T. 1901. 'Méthodes de calcul différentiel absolu et leurs applications', *Mathematische Annalen, 54*, 125–201.

Riemann, B. 1892. *Gesammelte mathematische Werke und wissenschaftlicher Nachlass*, 2nd ed. (ed. H. Weber), Leipzig: Teubner.

Renn, J., Sauer, T., Janssen, M., Norton, J. and Stachel, J. Forthcoming. *The genesis of general relativity*, 2 vols., Dordrecht, Boston and London: Kluwer.

Renn, J. and Sauer, T. 1999. 'Heuristics and mathematical representation in Einstein's search for a gravitational field equation', in [Goenner et alii, 1999], 87–125.

Sauer, T. 1999. 'The relativity of discovery. Hilbert's first note on the foundations of physics', *Archive for history of exact sciences, 53*, 529–575.

Scholz, E. (ed.) 2001. *Hermann Weyl's Raum-Zeit-Materie and a general introduction to his scientific work*, Basel, Boston and Berlin: Birkhäuser.

Schwarzschild, K. 1916. 'Über das Gravitationsfeld einer Kugel aus inkompressibler Flüssigkeit nach der Einsteinschen Feldtheorie', *Sitzungsberichte der Königlichen Preussischen Akademie der Wissenschaften* (*Berlin*), 424–434.

Stachel, J. 1995. 'History of relativity', in L.M. Brown, A. Pais and B. Pippard (eds.), *Twentieth century physics*, Philadelphia: Institute of Physics, vol. 1, 249–356.

Stachel, J. 2002. *Einstein from 'B' to 'Z'*, Boston, Basel and Berlin: Birkhäuser (Einstein Studies, vol. 9).

Weyl, H. 1918. *Raum·Zeit·Materie. Vorlesungen über allgemeine Relativitätstheorie*, Berlin: Springer. [Essential revisions in 3rd ed. 1919, 4th 1921 and 5th 1923.]

CHAPTER 64

D'ARCY WENTWORTH THOMPSON, *ON GROWTH AND FORM*, FIRST EDITION (1917)

T.J. Horder

This work has often been seen as the founding statement of an agenda for an analysis of biological phenomena in mathematical terms. Despite obscurities and stylistic idiosyncrasies, it is still referred to today because of its powerful arguments for the importance of physical forces in explaining the morphologies of organisms.

First publication. Cambridge: Cambridge University Press, 1917. 793 pages.

Second edition. 1942. 1116 pages.

Abridged version. Abridged and introduced by J.T. Bonner, 1961. [Repr. (introd. by S.J. Gould) as Canto paperback, 1992.]

Manuscript. None survives in his papers at the University of St Andrews, St Andrews, Scotland, but there are annotated copies of the book.

Translations of abridged version:

French. Forme et croisssance, Paris: Editions du Seuil, 1995. 3000 copies.

Italian. Crescita e forma, Milan: Bollati Boringhieri Editore, 1969, repr. 1992.

German. Wachstum und Form, Basel: Birkhäuser, 1973.

Greek. Athens: NTUA Press, 2001. 2000 copies.

Spanish. Sobre i crecimento y la forma, Madrid: H. Blume, 1981. [Repr. Madrid: Cambridge University Press Iberia, 2003. 1528 copies.]

Chinese. Sheng Chang He Xing Ta, Shanghai: Shanghai Scientific and Technical Publishers, 2003. 3000 copies.

Related articles: Maxwell (§44), Pearson (§56), Volterra (§73).

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

1 BIOGRAPHY

D'Arcy Wentworth Thompson (1860–1948) is remembered today for one published work, his monumental book *On growth and form*, first published in 1917 with a second edition in 1942. He was a biologist, but his inclusion in this volume on the history of mathematics is readily explained, because his attempt to extend the mathematical approach to cover the biological realm has been strongly influential up to the present time. The continuing impact of this book is evident in the fact that it is still in print. As we shall see this is all the more surprising given the many historically-based obscurities embedded in it. The precise nature of his influence is however far from clear.

Thompson was the only son of an eminent Edinburgh classics teacher [Thompson, 1958]. His mother's three brothers were biologists and medical men. After two years as a medical student at Edinburgh he reverted to his real vocation, that of a naturalist. He was trained in Cambridge, taking the Natural Science Tripos from 1880 to 1883, and there he was an exact contemporary of the future distinguished geneticists William Bateson and Walter Weldon. His friends included fellow undergraduates, the later physiologist C.S. Sherrington, the mathematician H.H. Turner, the philosopher W.R. Sorley, and A.N. Whitehead, mathematician later turned influential philosopher. At Cambridge, after initial courses in mathematics and physics, he was particularly influenced by Michael Foster, leader of the newly-emerging, rigorous, experimental approach to physiology—who also had an interest in embryology—and especially by Francis Balfour, the remarkably gifted embryologist who was pioneering a modern approach to comparative embryology, at the time very much a key discipline for biology generally. As a student Thompson translated Fritz Muller's *The fertilization of flowers*, for which he asked Charles Darwin to write a preface (one of his last writings). One suspects that Thompson might well have continued straightforwardly in the embryological direction, had not Balfour died suddenly in 1882, aged 31. His 'school' then fell apart.

At the age of 24 Thompson founded a new zoology department at Dundee. He later moved to St Andrews, where indeed he ended his long career, after 64 years as head of department. His interests ranged very wide, covering embryology, taxonomy, museum curating and collecting, fisheries and oceanography. He was equally at home with the classics; Aristotle was his particular favourite, about whom he often wrote. He produced scholarly glossaries of birds and fishes as referred to in the Classical Greek literature. It is not surprising that Stephen Gould described him as 'perhaps the greatest polymath this century'. However, apart from a steady stream of occasional papers, he published nothing of great scientific significance other than the book considered here, when he was in his late fifties. Its contents are summarised in Table 1.

2 STRUCTURE OF THE ARGUMENT

The book starts with the problem of scaling, that is, the direct consequences of the varying sizes of organisms, of surface to volume ratios, etc. The long Chapter 3 on 'The rate of growth' concerns growth curves, differential growth of anatomical parts and the effects on their form, i.e. shape, and its variations in different populations and under different conditions. The theme is now referred to as 'allometry'; it especially overlaps the

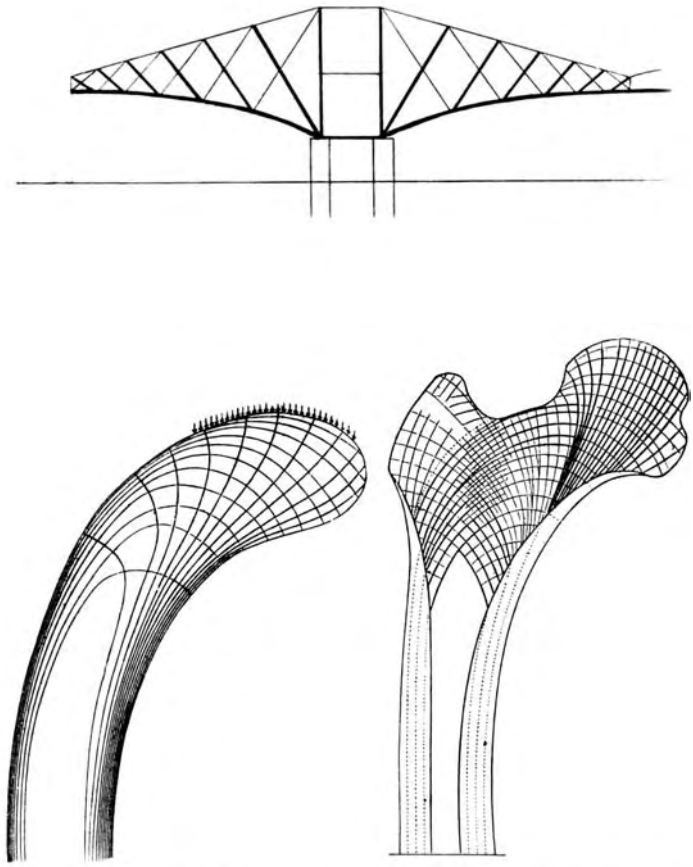
Table 1. Contents by Chapters of the book. The longer second edition is not different in layout or argument.

Ch.	Page	Topic
1	1	Introductory.
2	16	On magnitude.
3	50	The rate of growth.
4	156	On the internal form and structure of the cell.
5	201	The forms of cells.
6	277	A note on absorption.
7	293	The forms of tissues, or cell-aggregates.
8	346	The same.
9	411	On concretions, spicules, and spicular skeletons.
10	488	A parenthetic note on geodetics.
11	493	The logarithmic spiral.
12	587	The spiral shells of the foraminifera.
13	612	The shapes of horns, and of teeth or tusks: with a note on torsion.
14	635	On leaf-arrangement, or phyllotaxis.
15	652	On the shapes of eggs, and of certain other hollow structures.
16	670	On form and mechanical efficiency.
17	719	On the theory of transformations, or the comparison of related forms.
Epilogue	778–779	Index. [End 793.]

concerns of embryology. One reason for the length of this chapter is Thompson's interest in the mathematics of variation within populations, a controversial theme at the time among geneticists such as Galton, Weldon, Karl Pearson and Bateson (§56.2). Chapter 4 concerns the internal structure of cells; he interprets chromosome patterns and cell division as evidence for, and in terms of, mechanical forces such as surface tension. Chapter 5 deals in similar terms with the external shapes of different cell types. Chapter 6 on absorption addresses surface energy phenomena, which are seen as adjuncts to surface tension forces. Chapter 7 deals with the arranging of cells in groups (for example, honey combs). Chapter 8 considers the patterning of cells following cell division. Chapter 9 covers the role of mineral deposits in organismic structure, particularly shells and exoskeletons. Chapter 11 covers the specific case of helical patterns, especially in molluscan skeletons; here the mathematical description fits biological pattern especially closely and fruitfully.

The book now moves on to a more gross anatomical level; increasingly it merely draws the formal parallelism between aspects of shapes seen in anatomy as against structures found in the physical world (Figure 1). Chapter 14 reviews the Fibonacci series as a description of flower patterns.

Despite the vast diversity of issues and examples, the book is arranged in a logical and unified manner starting with the broadest principles, then moving from the simplest unit



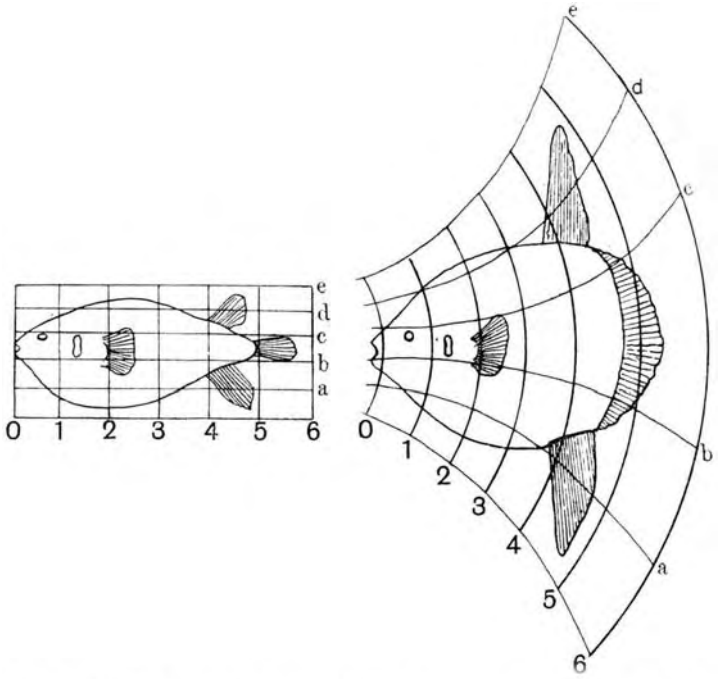
Figs 335, 345. The comparability of mechanical principles applicable to the Forth bridge, crane-head and femur.

Figure 1.

of biological structure, the cell, upwards in scale to biological structures and organs of increasing complexity.

3 THOMPSON'S KEY METHODOLOGICAL EXAMPLE

Undoubtedly the best remembered chapter in the book is the final one dealing with the theory of transformations. This is the culmination of the book and a synthesis of its overall message. It shows the relevance and application of Thompson's biomathematical approach to the problems of evolution of new species and phylogenesis. Here he tackles the largest scale morphological phenomena including whole organisms. The theory of transformations consists in showing how, after defining the shape of one structure (such as the species of



Figs 381, 382. The method of transformation as explanation for evolutionary transition between two fish species

Figure 2.

fish in Figure 2) in terms of an imposed Cartesian grid, a new shape (for example, a second fish species) can be derived through a coordinated geometric deformation of the grid.

The chapter stands out because of the methodology that Thompson introduces and the way that it is directly demonstrated in his striking diagrams. This is probably the most original part of the book. He sees this method as a way of avoiding traditional problems in taxonomy requiring the comparing of morphologies, such as population variations and the arbitrariness of the selection of specific individual anatomical features. His method deals with the *whole* pattern. But it is far from clear that the transformations are a pointer to actual mechanisms involved in biological evolution [Gould, 2002; Horder, 2002]. Are they any more than merely a descriptive device? The method certainly cannot explain evolutionary novelties of structure; but only relative morphological changes in arrangement of already existing structures in obviously comparable species.

4 HOW SHOULD SUCH A WORK BE APPROACHED TODAY?

4.1 The original intention of the author

The main problem in understanding and evaluating a historical work such as this one is that present perceptions are in danger of differing from those within which Thompson himself

would have seen his main achievement. Moreover, the way in which he happened to arrive at, and publish, his main objective—which is likely to reflect incidental circumstances and to bear the traces of his particular starting point and false trails on the way—probably needs to be separated from the finally achieved objective itself. Since we are dealing with a scientifically important and still influential work, we can hope to identify an important and enduring core message that is reasonably close to how Thompson would have understood it. The only valid starting point for a later analysis of a classic work must be an accurate assessment of what the work meant at the time and why it was seen as an important contribution to science. But in order to reach this point the historical fog between Thompson and ourselves has to be dispersed.

4.2 Language, style and presentation

In reading any historical work one faces a certain, inevitable foreignness of expression. The onus is on the present-day reader to make allowances for historical changes in style, and even in meaning of words. We also have to take into account the effects of publishing conventions of the time on the format of the work.

The modern reader is likely to find Thompson's style a dominant feature of the writing, particularly in the aspects of his 'classical' background which show up in his use of allusions and quotations (usually in their original languages) from a vast international range of historical sources. This can appear mannered and more literary than scientific. Nonetheless, the clarity of his meaning is rarely in doubt; the prose is often eloquent and forceful.

Thompson was by nature a naturalist and an encylopaedist; this explains the extraordinary breadth of his biological examples. And yet the book is a highly selective review of existing scientific literature. It is noticably thin on its coverage of experimental evidence, particularly in the field of embryology, which would seem to be most directly relevant to his main theme. Such a lack of coverage is even more glaringly obvious and unfortunate in the second edition; for example, he refers to Huxley [1932] only once despite the fact that J.S. Huxley had developed Thompson's methods in important ways. Although the book is most remembered for the chapter on transformation, its overall intention and approach is not really one of systematic reviewing of evidence or of presenting a new method. It is a compilation, verging on the anecdotal, intended to promote a certain, general view of biology. This explains why it cannot be described as a textbook, even less a work of general interest accessible to the non-experts; it is inevitably a 'one-off' and is perhaps best characterised as one extended essay. Although Thompson makes it quite clear how his approach relates to the major biological issues of the time (that is, evolution theory, genetics and phylogenesis), he is noticably brief in his treatment of them. The following passage from the epilogue (p. 778) sums up how he sees his presentation, and shows that he regarded the book as, in a sense, provisional and limited in ambition:

> In the beginning of this book I said that its scope and treatment were of so prefatory a kind that of other preface it had no need; and now, for the same reason, with no formal and elaborate conclusion do I bring it to a close. The fact that I set little store by certain postulates (often deemed to be fundamental) of our present-day biology the reader will have discovered and I have not

> endeavoured to conceal. But it is not for the sake of polemical argument that I have written, and the doctrines which I do not subscribe to I have only spoken of by the way. My task is finished if I have been able to shew that a certain mathematical aspect of morphology, to which as yet the morphologist gives little heed, is interwoven with his problems, complementary to his descriptive task, and helpful, nay essential, to his proper study and comprehension of Form.

4.3 Effects of historical context

No author can avoid being the product of his times. Failure to take historical context into account can cause much puzzlement to the modern reader; recognition of the context often provides the key that explains otherwise incomprehensible features. Again it is the duty of the reader to take account of the targets, fads, assumptions, debates or conflicts, and misunderstandings current at the time, if the work is to be properly judged and evaluated.

Although it is perfectly clear that he understood the biological context and implications of his contribution, Thompson's brevity and avoidance of engagement with the fundamental, polemical issues in biology of his time make it difficult to locate him within the scientific community. He was writing at a critical time for biology: Darwinism (including gradualism and adaptationism) was still being openly questioned; even the most basic aspects of the ultrastructural and macromolecular nature of cells were only beginning to be understood; Ernst Haeckel's notion that ontogeny recapitulates phylogeny still provided a sufficient rational explanation for morphology and as yet genetics provided few clues about ancestry or evolution. Thompson argues against various versions of 'final cause' or teleology, and in favour (like Wilhelm His) of explanations in the 'here and now'. Like Bateson [Coleman, 1970], Thompson was sceptical about new ideas on the role of chromosomes, partly because he argued against excessive reductionism (following Clerk Maxwell: compare §44). Below a certain size structure was irrelevant to biological phenomena. Not surprisingly he hardly mentions Gregor Mendel, August Weismann, Bateson or 'heredity'. A primary target for his attack was vitalism, a view of living system then still widespread. Thompson put his trust in method; like physiology, morphology needed to adopt the strictly explicit, objective and systematic methods that had been so successful in physics.

Throughout the text one can find indications of specific influences on Thompson's thought, particularly in his choice of authors and supporting evidence. He considered 'force' to be the most fundamental of causal factors; chemistry as well as biology could be reduced to it (pp. 1, 11). J.C. Maxwell was a major influence; his name is among the most frequently cited in the index. Vibrational models were potent images suggestive, and even illustrative, of mechanisms underlying such diverse processes as cell division, segmentation of bodily organisation or the striping of colour patterns in zebras or butterflies. The background assumptions and perspectives under which Thompson was operating can be inferred from the many remarkable similarities of his approach to that of Bateson, particularly in his volume *Problems of genetics* [1913], written at the same time. Interestingly, Bateson was not sympathetic to the mathematical approach, and yet he used many of the same analogies and physical models as Thompson [Coleman, 1970].

4.4 Subsequent perceptions and influence

It is clear that Thompson's original objective was to establish the general case for explanations of biological phenomena according to principles derived from physics and mathematics. This target, and some of his assumptions as he was writing the book, is probably most explicitly spelled out in early articles in which his ideas first emerged [Thompson, 1911, 1916]. His 1911 paper shows the extent to which he saw physiology as having already become a rigorous physics-based science; against this ideal morphology seemed sorely deficient methodologically. In a recent review of the history of biomathematics Keller [2002] has helped to put Thompson into context. Ball [1999] illustrates beautifully how he is still a key figure in some genres of current thinking about biological structure.

With a book such as this, one that—unusually—still has a continuing influence rather than being just an important but essentially archival historical document, one has to separate an evaluation of it as a work in its time, from the effects of its impact on succeeding generations. Its original impact may remain somewhat unclear, but succeeding generations may well have used it for their own, varying purposes.

It happens that a remarkable succession of future scientific opinion-formers took D'Arcy Thompson as an early model for their work. Following the first edition of *On growth and form* Huxley [1932] tried to refine the transformation method, both numerically and in algebraic form. For his more systematic and focused approach he coined the term 'allometry'. More recently Gould [1977] followed a similar path and attempted further modern refinements. Both dedicated their volumes to Thompson. Other prominent scientists much influenced by him include R.B. Goldschmidt, P.B. Medawar (1986), J.T. Bonner and Alan Turing. Turing attempted a directly mathematical solution to some of the most fundamental problems in embryonic development through his notion of the 'morphogen' and his 'reaction–diffusion model' [Hodges, 1983, ch. 8].

By the time (1942) of the second edition of *On growth and form*, Huxley, in marked contrast to Thompson, had moved on and had led the direction of advance in biology that was ultimately to prove most successful; in introducing 'The modern evolutionary synthesis' he completed the integration of modern genetics with the traditional concerns of evolutionary biology.

It seems possible that the enduring repute of Thompson's book is largely a function of the appeal it had to a particular succession to outstanding scientific communicators, who admired it for its style and erudition as much as for its science, and who, as young pioneers themselves, also found in Thompson a model, who appealed as a personality, especially as the maverick 'courageous loner'. In a revealing commentary on the book, Gould [2002] confirms the perhaps illusory qualities that have sustained its reputation and given the book its almost mythic status.

5 EVALUATION

It is difficult to evaluate this unorthodox volume, whose fame does not rest simply on an original and foundational scientific finding, a new theory or even a new methodology—at every point its originality could be questioned—but is inseparable from its style and

presentation. Thompson's claim to greatness does not rest on mathematical originality; he denied that he had special mathematical skills. The comment has often been made that his ideas are inapplicable in practice; even later refined versions like allometry are of limited usefulness in biology as a whole.

It has often been pointed out against Thompson that he was 'wrong' in almost all his views about biological fundamentals. As it turned out, his sceptical view of gradualistic Darwinian evolution or the phylogenetic and genetic basis of existing organic forms was unjustified. However, this judgement is unfair, given the state of flux of such issues at the time of the first edition. Thompson was not as much of an iconoclast as he has been taken to be, but he was clearly a free spirit, as well as being an opinionated and uninhibited thinker. He was also aware of the dangers he faced in taking these views and in promoting such a starkly alternative approach; he refers feelingly to the contempt with which his earlier mechanical views of embryogenesis had been met (p. 56). Undoubtedly one of his main motivations was to promote rigorous methods in biology, in the face of vitalism and the speculative thinking common at the time.

In the end we face a paradox. Thompson's specific intentions have not been fulfilled and his project could be said largely to have failed. The limitations of his approach in detail have become ever more obvious; biology cannot be reduced to mathematics, and now has to look to molecular biology for its foundations. And yet his abiding importance lies in his emphasis on morphology and the problem of explaining higher-level biological phenomena. There is no more eloquent statement of the limitations of reductionism available. In any scientific position that survives as a living force the position has to evolve along with changing historical contexts. Seen in this light Thompson's vision survives and may yet become even more important.

ACKNOWLEDGEMENT

Figures 1 and 2 are reproduced by permission of Cambridge University Press.

BIBLIOGRAPHY

Ball, P. 1999. *The self-made tapestry. Pattern formation in Nature*, Oxford: Oxford University Press.

Bateson, W. 1913. *Problems of genetics*, New Haven: Yale University Press.

Coleman, W. 1970. 'Bateson and chromosomes: conservative thought in science', *Centaurus*, *15*, 229–314.

Gould, S.J. 1977. *Ontogeny and phylogeny*, Cambridge, MA: Belknap Press.

Gould, S.J. 2002. *The structure of evolutionary theory*, Cambridge, MA: Belknap Press.

Hodges, A. 1983. *Alan Turing. The enigma*, London: Hutchinson.

Horder, T.J. 2002. 'Thompsonian transformations', in *Encyclopedia of life sciences*, vol. 18, London: Macmillan, 260–264.

Huxley, J.S. 1932. *Problems of relative growth*, London: Methuen.

Keller, E.F. 2002. *Making sense of life*: *explaining biological development with models, metaphors, and machines*, London: Harvard University Press.

Medawar, P.B. 1986. *Memoir of a thinking radish. An autobiography*, Oxford: Oxford University Press.

Thompson, D'Arcy W. 1911. 'Magnalia Naturae: or, The greater problems of biology', *Report of the British Association for the Advancement of Science*, 395–404.

Thompson, D'Arcy W. 1916. 'Morphology and mathematics', *Transactions of the Royal Society of Edinburgh*, *50*, 857–895.

Thompson, R. D'Arcy 1958. *D'Arcy Wentworth Thompson. The scholar-naturalist 1860–1948*, London: Oxford University Press.

CHAPTER 65

LEONARD DICKSON, *HISTORY OF THE THEORY OF NUMBERS* (1919–1923)

Della D. Fenster

Dickson provided an encyclopedic account of the history of number theory up to 1918. However, he omitted the important topic of quadratic reciprocity.

First publication. Volume I, *Divisibility and primality*, 1919. 486 pages. Volume II, *Diophantine analysis*, 1920. 803 pages. Volume III, *Quadratic and higher forms*, 1923. 313 pages. All vols. Washington: The Carnegie Institution.

Photoreprints. New York: Chelsea, 1966. Providence: American Mathematical Society 1999, 2002.

Related articles: Gauss on number theory (§22), Dirichlet (§37), Hilbert on number theory (§54).

1 BRIEF BIOGRAPHY OF THE AUTHOR

Born in Independence, Iowa in 1874, Leonard Eugene Dickson spent his boyhood in Cleburne, Texas and ultimately attended the University of Texas for his undergraduate and master's education [Albert, 1955; University of Texas Archives, 1899, 1914]. With his master's degree in hand and two years of teaching experience under his belt, Dickson chose as the place to pursue his doctorate the strong triumvirate of Eliakim Hastings Moore (1862–1932), Oskar Bolza, and Heinrich Maschke at the young University of Chicago over the up-and-coming Harvard with William Fogg Osgood and Maxime Bôcher. Dickson's mathematical career would ultimately hinge on this decision [Fenster, 1997, 9–13].

At the time, Chicago, with its sights set on emulating the German tradition of scholarship, stood in marked contrast to most American institutions. Specifically, Moore, Bolza and Maschke formed the core of the original far-sighted Chicago Mathematics Department, which promoted *both* research and teaching and which emphasized in its graduate program the training of future productive researchers [Parshall and Rowe, 1994, 363–426; Fenster, 1997, 10–11].

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

Moore, then drawn to group theory, inspired Dickson to write a thesis on (what we would call) permutation groups; this was duly done between 1894 and 1896 [Dickson, 1897]. Although group theory would remain among Dickson's research interests throughout his career, he would add finite field theory, invariant theory, the theory of algebras and number theory to his repertoire [Albert, 1955, 334–345 contains a bibliography of his work]. He reflected Chicago's influence—particularly that of Moore—in more ways than in his research interests, however. The department's sustained commitment to research, high standards for publication, and their vision for the American (as opposed to New England) mathematical community came to permeate Dickson's mathematical persona in these formative years. In the spring of 1900, the Chicago Mathematics Department invited him to join them as an assistant professor. From this position he made significant contributions to the consolidation and growth of the algebraic tradition in America [Fenster, 1997, 21]. Specifically, he spent 40 years (all but the first two) of his professional career on the faculty at Chicago, where he directed 67 Ph.D. students, wrote more than 300 publications, served as editor of the *Monthly* and the *Transactions* of the American Mathematical Society, and guided the Society as its President from 1916 to 1918.

2 WHY DICKSON MAY HAVE WRITTEN HIS *HISTORY OF THE THEORY OF NUMBERS*

Yet this mathematical workhorse, who played billiards and bridge by day and did mathematics from 8:30 p.m. to 1:30 a.m. every night [Albers and Alexanderson, 1991, 377], spent nearly a decade of his career writing a three-volume, 1602-page history of the theory of numbers. The lurking question is: why? As he explained it himself, he undertook this project because 'it fitted with my conviction that every person should aim to perform at some time in his life some serious useful work for which it is highly improbable that there will be any reward whatever other than his satisfaction therefrom' (vol. 2, xxi). Although he viewed it as 'highly improbable', this altruistic mission paid handsome rewards for Dickson as this historical study ultimately led to his celebrated work in the arithmetics of algebras [Fenster, 1998].

Dickson's most distinguished student, A. Adrian Albert, has suggested that Dickson wrote the book to become more acquainted with number theory: 'Dickson always said that mathematics is the queen of sciences, and that the theory of numbers is the queen of mathematics. He also stated that he had always wished to work in the theory of numbers and that he wrote his monumental *History of the Theory of Numbers* so that he could know all of the work which had been done in the subject' [Albert, 1955, 333].

Dickson's developing research interests substantiate this claim. Of the 200 papers he wrote prior to 1923, the year he published the third (and final) volume of his *History*, only ten considered number-theoretic topics. In 1927, however, his pure mathematical researches began to focus on additive number theory, and in particular on the ideal Waring theorem. In a long series of papers he and his students provided an almost complete verification of the theorem that loosely states that every positive integer is a sum of I integral n-th powers for sufficiently large I. Dickson also guided 29 of his last 32 doctoral students in number-theoretic dissertations [University of Chicago Archives, 1931, 1938, 1941]. These

29 students, along with Dickson's contributions to the ideal Waring theorem and three texts on number theory that he published as [Dickson, 1929, 1930, 1939] seem to indicate that if he intended for his historical study to acquaint him with the subject so that he could work in the field himself, he had certainly accomplished what he set out to do.

However, Dickson may have undertaken this historical work for more nationalistic reasons. For example, he initially sought out the Carnegie Institution of Washington, one of the new national agencies created to promote what we now call basic research [Reingold and Reingold, 1981, 7], as a possible publisher of the project. From his perspective, as he wrote to Carnegie President R.S. Woodward, '[i]t would seem desirable to have undertaken in this country something of the kind done by the British Association, the Deutsche Mathematiker Vereinigung, etc., in the preparation by specialists of note of extensive Reports each covering an important branch of science' [Dickson to Woodward, 1911]. After describing his 'ideal of a mathematical report' that would appeal both to specialists and non-specialists, Dickson admitted that he had 'already given a solid year's work to such an expository Report on the theory of numbers (integral and algebraic)' (*ibidem*). Thus the British and German 'mathematical Report[s]', and, in particular, the lack of similar offerings in America, may have encouraged Dickson to write his own compendium on the subject of number theory. In the case of graduate training, it was not at all unusual for the American mathematicians to look to the Europeans for ideas [Parshall and Rowe, 1994]. The initiative that Dickson outlined in his letter to Woodward, however, required not only an acquaintance with the European literature but also an awareness of a perceived void in American publications. Moreover, the opening sentence of his letter to Woodward seems to suggest that he wanted to raise American mathematics to the European standard in this particular realm. Throughout his career, he remained avidly committed to establishing standards of excellence for and in the community of *American* mathematicians [Fenster, 1998, 1999].

3 THE STYLE AND CONTENT OF DICKSON'S *HISTORY OF THE THEORY OF NUMBERS*

Table 1 summarises the contents of Dickson's *History.* His view of the role of the historian dictated how he both prepared and wrote his book. As he saw it, '[w]hat is generally wanted [in a historical study] is a full and correct statement of the facts, not an historian's personal explanation of those facts. The more completely the historian remains in the background, the better the history. Before writing such a history, he must have made a more thorough search for all the facts than is necessary for the conventional history' (vol. 1, xx). For him this 'thorough' search required a trip overseas to visit European libraries and collect various number theoretic references. The University of Chicago, apparently, supported this type of international research travel since they granted Dickson a leave of absence. For the necessary funds, he sought travel support for his research from the Carnegie Institution of Washington. From a purely pragmatic perspective, his *History* confirms the importance of recent 'technical innovations'—such as the railroad, steamship, and telegraph—in the internationalization of science [Parshall, 1996, 293; Lehto, 1998, 1–2]. Specifically, Dickson could not have undertaken, much less completed, his *History* without the recent advances of the railroad to take him to the East Coast of the United States, the steamship to carry

Table 1. Contents by chapters of Dickson's *History of the theory of numbers*.

Ch.	Page	Topics
Volume I. *Divisibility and primality.*		
1	3	Perfect, multiply perfect, and amicable numbers.
2	51	Formulas for the number and sum of divisors, problems of Fermat and Wallis.
3	59	Fermat's and Wilson's theorems, generalizations and converses; symmetric functions of $1, 2, \ldots, p-1$, modulo p.
4	105	Residue of $(u^{p-1}-1)/p$ modulo p.
5	113	Euler's f-function, generalizations; Farey series.
6	159	Periodic decimal fractions; periodic fractions; factors of $10^n \pm 1$.
7	181	Primitive roots, exponents, indices, binomial congruences.
8	223	Higher congruences.
9	263	Divisibility of factorials and multinomial coefficients.
10	279	Sum and number of divisors.
11	327	Miscellaneous theorems on divisibility, greatest common divisor, least common multiple.
12	337	Criteria for divisibility by a given number.
13	347	Factor tables, lists of primes.
14	357	Methods of factoring.
15	375	Fermat numbers $F_n = 2^{2n}+1$.
16	381	Factors of $a^n \neq b^n$.
17	393	Recurring series; Lucas' u_n, v_n.
18	413	Theory of prime numbers.
19	441	Inversion of functions; Möbius's function $\mu(n)$; numerical integrals and derivatives.
20	453	Properties of the digits of numbers.
	467	Author index. 484–486 Subject index.
Volume II. *Diophantine analysis*.		
1	1	Polygonal, pyramidal and figurate numbers.
2	41	Linear Diophantine equations and congruences.
3	101	Partitions.
4	165	Rational right triangles.
5	191	Triangles, quadrilaterals, and tetrahedra.
6	225	Sum of two squares.
7	259	Sum of three squares.
8	275	Sum of four squares.
9	305	Sum of n squares.
10	325	Number of solutions of quadratic congruences in n unknowns.

Table 1. (*Continued*)

Ch.	Page	Topics
11	329	Liouville's series of eighteen articles.
12	341	Pell equation; $ax^2 + bx + c$ made a square.
13	401	Further single equations of the second degree.
14	435	Squares in arithmetical or geometrical progression.
15	443	Two or more linear functions made squares.
16	459	Two quadratic functions of one or two unknowns made squares.
17	485	Systems of two equations of degree two.
18	491	Three or more quadratic functions of one or two unknowns made squares.
19	497	Systems of three or more equations of degree two in three or more unknowns.
20	533	Quadratic form made an nth power.
21	545	Equations of degree three.
22	615	Equations of degree four.
23	673	Equations of degree n.
24	705	Sets of integers with equal sums of like powers.
25	717	Waring's problem and related results.
26	731	Fermat's last theorem, $ax^r + by^s = cz^t$, and the congruence $x^n + y^n = z^n \pmod p$.
	777	Author index. 799–803 Subject index.
Volume III. *Quadratic and higher forms.*		
1	1	Reduction and equivalence of binary quadratic forms, representation of integers.
2	55	Explicit values of x, y in $x^2 + Dy^2 = g$.
3	60	Composition of binary quadratic forms.
4	80	Orders and genera; their composition.
5	89	Irregular determinants.
6	92	Number of classes of binary quadratic forms with integral coefficients.
7	198	Binary quadratic forms whose coefficients are complex integers or integers of a field.
8	203	Number of classes of binary quadratic forms with complex integral coefficients.
9	206	Ternary quadratic forms.
10	225	Quaternary quadratic forms.
11	234	Quadratic forms in n variables.
12	253	Binary cubic forms.
13	259	Cubic forms in three or more variables.
14	262	Forms of degree $n \geqslant 4$.

Table 1. (*Continued*)

Ch.	Page	Topics
15	269	Binary Hermitian forms.
16	279	Hermitian forms in n variables and their conjugates.
17	284	Bilinear forms, matrices, linear substitutions.
18	289	Representation by polynomials modulo p.
19	293	Congruencial theory of forms.
	303 Author index. 309–313 Subject index.	

him across the Atlantic and the telegraph to aid him with his correspondence. The time was right for such a comprehensive undertaking.

Dickson's description of this historical undertaking as 'serious useful work', however, proved more than accurate. This was no hastily written history of number theory. On the contrary, he had planned both the content of his project and the precise method that he would follow to present the details of his study. He revealed the scope of his plans when he explicitly stated his bold intention to 'give an adequate account of the entire literature of the theory of numbers' (vol. 1, iii). As for his method, the following excerpt reveals both the thoroughness of his study and the historiographic view that he maintained throughout this work. As for the presentation, a typical page (vol. 1, 5) reveals the stylistic manifestation of his historiographic view:

> Hrotsvitha, a nun in Saxony, in the second half of the tenth century, mentioned the perfect numbers 6, 28, 496, 8128.
> Abraham Ibn Ezra (1167), in his commentary to the Pentateuch, Ex. 3, 15, stated that there is only one perfect number between any two successive powers of 10.
> Rabbi Josef b. Jehuda Ankin, at the end of the twelfth century, recommended the study of perfect numbers in the program of education laid out in his book 'Healing of Souls.'
> Jordanus Nemorarius (1236) stated (in Book VII, props. 55, 56) that every multiple of a perfect or abundant number is abundant, and every divisor of a perfect number is deficient. He attempted to prove (VII, 57) the erroneous statement that all abundant numbers are even.
> Leonardo Pisano, or Fibonacci, cited in his Liber Abbaci of 1202, revised about 1228, the perfect numbers
>
> $$\frac{1}{2}2^2(2^2-1)=6, \qquad \frac{1}{2}2^3(2^3-1)=28, \qquad \frac{1}{2}2^5(2^5-1)=496,$$
>
> excluding the exponent 4 since 2^4-1 is not prime. He stated that by proceeding so, you can find an infinitude of perfect numbers.

In 1602 pages, Dickson never swerved from this comprehensive, facts-only style of writing: Hrotsvitha mentioned, Ezra stated, Rabbi Josef recommended, Nemorarius stated,

Fibonacci cited, etc. This strict style, in the opinion of the number-theorist D.N. Lehmer, made 'the book [...] not so much a history as a list of references from which a history of the theory of numbers might be written' [Lehmer, 1919–1920, 131–132].

In some cases, however, the *sum total* of the facts departed from the strictly internalistic (to use a modern historiographic adjective) style followed by Dickson, and revealed a much broader view of the theory of numbers. As we saw above, he included a 12th-century rabbi as a contributor to the development of perfect numbers and described his contribution as one who 'recommended the study of perfect numbers in the program of education laid out in his book "*Healing of Souls*" '. The preceding page included more 'facts' on the ethical importance of perfect numbers (vol. 1, 4):

> Iamblichus (about 283–330) [...] remarked that the Pythagoreans called the perfect number 6 marriage, and also health and beauty (on account of the integrity of its part and the agreement existing in it).
> Aurelius Augustinus (354–430) remarked that, 6 being the first perfect number, God effected the creation in 6 days rather than at once, since the perfection of the work is signified by the number 6. [...].
> Alcuin (735–804), of York and Tour, explained the occurrence of the number 6 in the creation of the universe on the ground that 6 is a perfect number. The second origin of the human race arose from the deficient number 8; indeed, in Noah's ark there were 8 souls from which sprung the entire human race, showing that the second origin was more imperfect than the first, which was made according to the number 6.

Hence, as Lehmer pointed out in his review of this volume for the *Bulletin of the American Mathematical Society*, 'one is struck in glancing through the book by the remarkable combination of superstition, fancy, scientific curiosity, and patient, plodding experiment that has figured in advancing the science of the theory of numbers' [Lehmer, 1919–1920, 125]. Dickson may or may not have minded this sort of comment made *about* his book, but he certainly would have never drawn the conclusion in the book itself.

4 ONE SALIENT OMISSION

Dickson's purportedly complete history of the theory of numbers lacks the quintessential topic of elementary number theory, the law of quadratic reciprocity. This law relates the solvability of the congruences $x^2 \equiv p \pmod{q}$ and $x^2 \equiv q \pmod{p}$ for p and q distinct, odd primes. Specifically, if p or q is of the form $4k + 1$ (for $k \in \mathbf{Z}$), the two congruences are both solvable or both not solvable. If p and q are both of the form $4k + 3$ (for $k \in \mathbf{Z}$), one of the congruences is solvable and the other is not. In terms of the Legendre symbol, for p and q distinct, odd primes,

$$\left(\frac{p}{q}\right)\left(\frac{q}{p}\right) = (-1)^{(\frac{p-1}{2})(\frac{q-1}{2})}. \tag{1}$$

This law, as Dickson described it himself, 'is doubtless the most important tool in the theory of numbers and occupies the central position in its history. Its generalizations form

a leading topic, past and present, in the theory of algebraic numbers' [Dickson, 1929, 30]. Given that the development of algebraic number theory grew, in large part, out of efforts to generalize quadratic reciprocity, it seems all the more unusual that a supposedly comprehensive *History of the theory of numbers* included no discussion of this area.

Why, then, did Dickson exclude the account of 'this most important' tool from his *History*? The historical record suggests that Dickson did not intend for this omission to occur. In his closing remarks in the preface to volume II (written in April 1920), Dickson refers to a Volume III as the 'concluding' volume in the series (vol. 2, xii). In April 1921 he wrote to President John C. Merriam of the Carnegie Institution regarding the 'final (third) volume' of his *History*. In particular, he outlined the seven chapters of this third volume as '1. Quadratic residues; 2. Quadratic reciprocity law; 3. Higher residues and reciprocity laws; 4. Binary quadratic forms; 5. Class number of [quadratic forms]; 6. Quadratic forms in 3 or more variables; 7. Higher Forms' [Dickson to Merriam, 1921].

Volume III appeared in 1923, 'promptly' prepared, as Dickson described it in the preface, 'owing to the favorable reception accorded to the first two volumes of this history' (vol. 3, iii). Early in the text of this third volume, nestled in his history of binary quadratic forms, Dickson points us forward to a *fourth* volume (vol. 3, 3). In this parenthetical remark, he indicated his plan to include the quadratic reciprocity law in the fourth volume. But, of course, as we know now, it never appeared. What happened to it?

The answer involves Albert Everett Cooper, a University of Chicago graduate student from 1924 to 1926. Cooper earned his Ph.D. in mathematics in the spring of 1926 under Dickson's guidance with his historical dissertation, 'A topical history of the theory of quadratic residues' [Cooper, 1926]. He wrote this dissertation with the intention that it appear as a chapter in the fourth volume of Dickson's *History*, which would contain a separate chapter on the history of quadratic reciprocity.

Originally, the Carnegie Institution of Washington agreed to issue the 'fourth and final volume' of Dickson's *History*. But Dickson had other ideas. He proposed that the Carnegie Institution no longer plan to publish the fourth volume, but instead publish one of his two new forthcoming treatises on number theory. Dickson attempted to secure publication for the fourth volume elsewhere, but with no success. Although Cooper prepared the manuscript for the Carnegie Institution in 1929, they did not publish it.

The A.E. Cooper Papers in the University of Texas Archives house the scores of papers, notes, and communiqués of various forms exchanged between Cooper and Dickson on the history of quadratic reciprocity. The organized and polished pieces of this collection seem to represent the page proofs of a book (the fourth volume?) written in the same spirit and style of the first three volumes of Dickson's *History*.

5 RECEPTION OF DICKSON'S *HISTORY OF THE THEORY OF NUMBERS*

The reviews of this masterpiece suggest that Dickson accomplished this historical endeavor with the same prowess as his work in pure mathematics. As Robert Carmichael, a number theorist who read the proof sheets for the entire second and third volumes, expressed it in his review for the *Monthly* [Carmichael, 1919, 397, 403]:

> To give an adequate account of the entire literature of so vast a subject and one of such long history as the theory of numbers is an undertaking of enormous magnitude; and it is carried through in this work with a marvelous success in the presence of which one must pause in admiration. Henceforth this history will be indispensable to all investigators in the theory of numbers. [. . .] It is a piece of work for which one cannot find a parallel in the whole of scientific history.

Dickson's *History* remains the classic reference on number theory up to 1918. It provided—and provides?—an 'indispensable' source for those lacking adequate library facilities [Carmichael, 1919, 397]. In particular, as Dickson intended, the many 'amateurs' interested in mathematics benefited from this (reputably) comprehensive, available account of number theory (vol. 2, xx; [Carmichael, 1919, 397]). As for the professional mathematician, in his review of volume I Lehmer emphasized 'the greatest need for just such a piece of work to promote efficiency among the professional workers in this field and to prevent them from wasting their time on problems that have already been adequately treated, and also to suggest other problems which still defy analysis' [Lehmer, 1919–1920, 132]. The research mathematician would gain so much more than 'efficiency' by the time all three volumes appeared in print.

Dickson's 'systematic' study of Diophantine analysis for the second volume of his *History*, for example, provided him with a unique, sweeping perspective on this area of mathematics. From this vantage point, he could assert that '[s]ince there already exist too many papers on Diophantine Analysis which give only special solutions, it is hoped that all devotees of this subject will in future refrain from publication until they obtain general theorems on the problem attacked if not a complete solution of it. Only in this way will the subject be able to retain its proper position by the side of other virile branches of mathematics' (vol. 2, xx). Dickson, in no uncertain terms, made this assertion with authority. Who better than a prominent research mathematician studying the 'disjointed elements' of Diophantine analysis, could so confidently declare in essence that '[i]deas rather than computations are needed in this field' [Carmichael, 1921, 72–73]? Dickson's firm grasp on the past allowed him to see what would lead to a prosperous future for Diophantine analysis. Carmichael emphasized the value of such forecasting when he wrote that '[w]hen a master, with the work of the past well in mind, tries to see the trend of the future, his judgment will be a matter of interest whether or not the direction of progress turns out to be such as he anticipates. It may even throw some light on the difficult question as to the way in which new discoveries arise' [Carmichael, 1923, 262]. Interestingly, Dickson himself would devote the final 15 years of his mathematical career focused on establishing a general result in Diophantine analysis. Simply put, Dickson made the history of number theory work in very utilitarian ways—far beyond serving solely as a reference volume—for the research mathematician.

Moreover, in the early years of the 20th century, the *Proceedings of the London Mathematical Society* contains more references to Dickson's *History* than to any other historical source [Rice, 2002]. Thus other mathematicians regarded it as a historical resource for their mathematical research. This met the precise need described by Bashford Dean in his contribution to the discussion on the pages of *Science* regarding the best way the newly

formed Carnegie Institution of Washington could advance science in this country. Dean dogmatically described his view that '[a]ll workers in science need skilful and energetic help in the thankless drudgery of reference hunting' [Dean, 1902, 643].

Dickson's *History of the theory of numbers* not only inspired his contemporaries [Lehmer, 1919–1920; Smith to Merriam, 1921], but also the next generation of mathematicians. Richard Guy, for example, purchased Dickson's *History* when he was about 17 years old and found it 'better than getting the whole works of Shakespeare and heaven knows what else' [Albers and Alexanderson, 1993, 136]. In his very different approach to a history of number theory, Oystein Ore refers his readers to Dickson's three-volume *History* for 'a complete, encyclopedic account of the history of the discoveries in number theory up to 1918' [Ore, 1988, 359d]. With their consistent reference to Dickson's *History* 'for more details' [Hardy and Wright, 1938] testify further to the significance of Dickson's work. Even more contemporary, Alan Baker acknowledges Dickson's *History* as an 'excellent source' in his *A concise introduction to the theory of numbers* [Baker, 1984]. Then and now, in histories and mathematical studies of number theory, Dickson's *History* serves as the quintessential reference for number theory up to 1918.

Still in print today, Dickson's *History* may belong to the tiny collection of 'books on the history of mathematics that were written over fifty years ago [and] continue to attract readers today' [Rowe, 2001, 590]. David Eugene Smith hinted at this 80 years ago when he candidly wrote: '[n]othing has ever come out in this country on the history of mathematics that is so epoch-making as this work. It is, of course, much more than a mere history, because it contains the theory as well' [Smith to Merriam, 1921].

BIBLIOGRAPHY

Albers, D.J. and Alexanderson, G.L. 1991. 'A conversation with Ivan Niven', *College mathematics journal, 22*, 370–402.

Albers, D.J. and Alexanderson, G.L. 1993. 'A conversation with Richard K. Guy', *College mathematics journal, 24*, 123–148.

Albert, A.A. 1955. 'Leonard Eugene Dickson 1874–1954', *Bulletin of the American Mathematical Society, 61*, 331–345.

Baker, A. 1984. *A concise introduction to the theory of numbers*, Cambridge: Cambridge University Press.

Carmichael, R.D. 1919, 1921, 1923. Review of *History of the theory of numbers*, vols. 1, 2, 3, *American mathematical monthly, 26*, 396–403; *28*, 72–78; *30*, 259–262.

Cooper, A.E. 1926. 'A topical history of the theory of quadratic residues', University of Chicago Archives.

Dean, B. 1902. 'The Carnegie Institution', *Science, 16*, 641–644.

Dickson, L. to J.C. Merriam, 26 April 1921, Carnegie Institution Archives, Washington DC, Dickson Papers.

Dickson, L. to R.S. Woodward, 11 February 1911, *Ibidem.*

Dickson, L. 1897. 'The analytic representation of substitutions on a power of a prime number of letters with a discussion of the linear group', *Annals of mathematics, 11*, 65–143.

Dickson, L. 1929. *Introduction to the theory of numbers*, Chicago: The University of Chicago Press.

Dickson, L. 1930. *Studies in the theory of numbers*, Chicago: The University of Chicago Press (The University of Chicago Science Series).

Dickson, L. 1939. *Modern elementary theory of numbers*, Chicago: University of Chicago Press.
Fenster, D.D. 1998. 'Leonard Eugene Dickson and his work in the arithmetics of algebras', *Archive for history of exact sciences, 52*, 119–159.
Fenster, D.D. 1997. 'Role modeling in mathematics: the case of Leonard Eugene Dickson (1874–1954)', *Historia mathematica, 24*, 7–24.
Fenster, D.D. 1999. 'Leonard Dickson's *History of the theory of numbers*: an historical study with mathematical implications', *Revue d'histoire des mathématiques, 5*, 159–179.
Fenster, D.D. 2002. 'American initiatives toward internationalization: the case of Leonard Dickson', in K.V.H. Parshall and A. Rice (eds.), *Mathematics unbound. The evolution of an international mathematical community, 1800–1945*, Providence: American Mathematical Society; London: London Mathematical Society, 311–333.
Hardy, G.H. and Wright, E.M. 1938. *An introduction to the theory of numbers*, 1st ed., London: Oxford University Press. [Later eds. to 5th, 1979.]
Lehmer, D.N. 1919–1920. 'Dickson's History of the theory of numbers', *Bulletin of the American Mathematical Society 26*, 125–132.
Lehto, O. 1998. *Mathematics without borders: a history of the International Mathematical Union*, New York: Springer-Verlag.
Ore, O. 1948. *Number theory and its history*, New York: McGraw Hill. [Repr. New York: Dover, 1988.]
Parshall, K.H. 1996. 'How we got where we are: an international overview of mathematics in national contexts (1875–1900)', *Notices of the American Mathematical Society, 43*, 287–296.
Parshall, K.H. and Rowe, D. 1994. *The emergence of an American mathematical research community: J.J. Sylvester, Felix Klein, and E.H. Moore*, Providence: American Mathematical Society; London: London Mathematical Society.
Reingold, N. and Reingold, I. 1981. *Science in America: a documentary history, 1900–1939*, Chicago: University of Chicago Press.
Rice, A. 2002. Private communication.
Rowe, D.E. 2001. 'Looking back on a bestseller: Dirk Struik's *A concise history of mathematics*', *Notices of the American Mathematical Society, 6*, 590–592.
Smith, D.E. to J.C. Merriam, 24 May 1921, Carnegie Institution Archives, Washington, DC, Dickson Papers.
University of Chicago 1931. Department of Special Collections, 'Doctors of Philosophy, June 1893 – April 1931', in *Announcements: The University of Chicago*, vol. 31, no. 19 (15 May), Chicago: University of Chicago Press.
University of Chicago 1938. Department of Special Collections, *Register of Doctors of Philosophy 1938–1939*, vol. 38, no. 4 (25 May).
University of Chicago 1941. *Records of the University of Chicago Convocation Programs*, Convocations 191–207 (1938–1941).
University of Texas 1899. The Center for American History, University of Texas Memorabilia Collection, 1899. The University of Texas, vol. 1, no. 3, August.
University of Texas 1914. The Center for American History, University of Texas Memorabilia Collection, 1914. 'Who's who at Texas?' *The Alcalde*, January, 266–268.
University of Texas, The Center for American History, Archives of American Mathematics, A.E. Cooper Papers.

CHAPTER 66

PAUL URYSOHN AND KARL MENGER, PAPERS ON DIMENSION THEORY (1923–1926)

Tony Crilly

The papers of Urysohn and Menger provided a definition of 'local dimension' of a topological space and a substantial accompanying theory. This contribution suggested a fruitful research direction during a time of rapid development for topology.

Urysohn. 'Mémoire sur les multiplicités Cantoriennes' and '(suite)', *Fundamenta mathematicae, 7* (1925), 30–137, 378–380; and *8* (1926), 225–359. Dated 20 March 1923, Moscow.

Russian translation. In *Papers on topology and other branches of mathematics*, 2 vols. (ed. P.S. Alexandrov), Moscow and Leningrad: Gostekhizdat, 1951.

Menger. 'Über die Dimensionalität von Punktmengen, Erster Teil' and 'II. Teil', *Monatshefte für Mathematik und Physik, 33* (1923), 148–160; and *34* (1926), 137–161. Submitted 12 December 1923, published April 1924; and submitted 6 October 1924, published September 1926.

Related articles: Cantor (§46), Baire and Lebesgue (§59), Seifert/Threlfall and Hopf/Alexandrov (§76).

1 ANCESTRY

Grappling with the problem of giving a precise definition of dimension has been a continuing theme in mathematics since Euclid. Stated in modern terms, how can a number be assigned to a set of points that is invariant under a one-to-one bi-continuous (that is, topological) transformation of these sets? Furthermore, how can a *theory* of dimension be constructed? This was the challenge that attracted Paul Urysohn and Karl Menger at the beginning of their professional lives.

Fifty years before, Georg Cantor (1845–1918) had discovered results that run counter to the usual intuition for assigning dimension numbers to typical geometrical objects like

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

lines, squares, and cubes. He startled geometers by displaying a one-to-one correspondence between a 'one-dimensional' line and a 'two-dimensional' square. This fact put the accepted notion of dimension in jeopardy, but he saw that it might be saved. He attempted to show that by requiring *continuity* the paradox could be resolved, to prove that a Euclidean line and square could not be mapped into each other under a bi-continuous one-to-one transformation. Around the same time, other mathematicians attempted to prove this 'invariance of dimension', but without success. The general proof that such a topological transformation (a homeomorphism) between Euclidean spaces of dimensions m, n existed if and only if $m = n$ turned out to be elusive until L.E.J. Brouwer (1881–1966) gave two proofs, one in 1911 and another in 1913.

Obtaining a clear definition of dimension that would serve for *arbitrary* sets of points, and at the same time agree with intuitive notions remained a problem. What was meant by a curve and a surface posed a concomitant question and various answers presented paradoxical results. The definition of a curve due to Camille Jordan (1838–1922), of being the image of a continuous mapping of a closed interval, yielded results at variance with intuitive beliefs of what dimension should be. With Jordan's definition, the 'two-dimensional' solid square could be regarded as a curve, whereas sets of points consisting of 'one-dimensional' line segments could be produced which were not curves in the Jordan sense. Could an adequate notion of dimension be used to separate such examples as these?

In the first decade of the 20th century, several mathematicians attempted to define dimension, notably, Brouwer and Henri Poincaré (1854–1912). Poincaré's scheme of using 'cuts' was not immune to curious results: the dimension of the double cone, ostensibly two-dimensional, turned out to be one-dimensional. Brouwer's idea, which he termed the *integral dimension*, was to say that a continuum (a closed connected set of points) is called 'n-dimensional' if it can be divided into separate pieces by means of one or more $(n - 1)$-dimensional continua. In modern terms this is referred to as the large inductive dimension and is denoted by *Ind*. Both his and Poincaré's recursively framed definitions were global in that they spoke of the dimension of a whole space.

Though these ideas held promise, neither mathematician was primarily concerned with constructing a formal mathematical theory at this stage. To add to the interest, Henri Lebesgue (1875–1941) put forward the quite different notion of 'covering dimension' (denoted by *dim*), which he only developed rigorously in the 1920s. In Lebesgue's definition, if each point of a domain D belongs to at least one of a finite number of closed sets, and if these sets have a sufficiently small diameter, then there are points common to at least $n + 1$ of these sets.

These definitions of dimension briefly discussed are topological in character, in that they can be freed from metric considerations. Yet another concept of dimension, but one wholly dependent on metric considerations, is due to Felix Hausdorff (1868–1942). This was articulated by him in 1919 when he started from a generalization of Lebesgue measure due to Constantine Carathéodory (1873–1950). Hausdorff gave a measure-theoretic characterization for Euclidean spaces, leading to what is now known as the Hausdorff dimension.

The origins of Hausdorff dimension can be traced to Weierstrassian investigations of non-differentiable continuous curves that took place in the 1870s. The peculiarity of Hausdorff dimension is that it need not be integral, as for example in the case of the Koch

'snowflake curve' that has Hausdorff dimension $(\ln 4/\ln 3) = 1.2618\ldots$, the calculation being based on length. Unaware of Hausdorff's work, Georges Bouligand (1889–1979) recreated this theory of dimension during the 1920s. In recent times it has gained prominence through the theory of fractals. There is a connection between Urysohn and Menger's theory, but as Hausdorff dimension is not a topological concept, it will not be pursued here. For detailed historical information on the origins and initial development of dimension theory as a whole, see [Johnson, 1979, 1981; Crilly with Johnson, 1999].

2 TWO PHYSICISTS

Paul Urysohn (Pavel Samuilovich Uryson) (1898–1924) was born of a Jewish family in Odessa, on the Black Sea, where his father was a wealthy financier. He had a solitary childhood and from an early age was involved with academic study, chemistry and physics being his favourite subjects. Life changed in 1909 when his mother died, and from then on he was watched over by Lina, the youngest of three much older sisters. The following year the family moved to Moscow where Paul was sent to a gymnasium school. He wanted to become a physicist and, quite remarkably, while still a schoolboy, began work at the Shanyavskii University under the supervision of P.P. Lazarev (1878–1942). His success as a physical scientist seemed assured as his experimental work in X-rays led to a published paper on Coolidge Tube radiation [Arkhangelskii and Tikhomirov, 1998; Cameron, 1982].

Karl Menger (1902–1985) came from a background of academia, the arts, and public service. His father was the Austrian economist, Carl Menger (1840–1921), a professor at the University of Vienna, and his mother, Hermione Andermann, was a successful novelist. One of his uncles was Anton Menger (1841–1906), a noted social scientist and also a professor at the university, and another was a long-standing member of the Austrian Reichstag. At the Döblinger gymnasium school, Karl was in the company of future Nobel prizewinners, Wolfgang Pauli (1900–1958) and Richard Kuhn (1900–1967). With such an academic pedigree and surrounded by academic brilliance, he entered the University of Vienna during the autumn of 1920, with an imperative to do well. The appointment of Hans Thirring had added lustre to an excellent physics department, and young Menger's plan was to study theoretical physics.

3 PAUL URYSOHN, MATHEMATICIAN

In 1915 Urysohn entered Moscow University but changed his field to mathematics. He attended the lectures of Dimitrii F. Egorov (1869–1931) and of Egorov's student Nikolai N. Luzin (1883–1950). A man of the old school, Egorov was formal in his approach to students, while Luzin was relaxed and could enter into their lives. He gathered around him a group of 'Luzitanians' and Urysohn was a leading member. After completing a rigorous mathematical training, and graduating in 1919, Urysohn continued with post-graduate work. He met Paul (Pavel Sergeevich) Alexandrov (1896–1982) and the two became great friends, known around the university as the 'two P.S's'. Under Luzin's influence, Urysohn prepared his doctoral thesis in the spring of 1921.

Urysohn's doctorate was completed by June 1921 for a thesis on integral equations. This was credited with founding non-linear analysis in Russia, and around the same time, he

produced outstanding work on convex and differential geometry. He became a member of staff of the Institute of Mathematics and Mechanics, and a professor at the second Moscow University. In the summer, Egorov suggested he might attempt to formulate an *intrinsic* definition of curves and surfaces, intrinsic because they were to be independent of the containing space. Urysohn's task was to describe the most general point sets that merit being called lines and surfaces, and this quest led him to seek a rigorous definition of dimension.

In August Urysohn was on holiday with other Luzitanians, renting a dacha near Bolshev on the Kalyazmy river. Topology in Russia was given an impetus by this group and he was at the centre of this movement. It was an exciting period in his life, with a lost childhood regained and his adventurous spirit overcoming the more morose side of his character. In the Russian countryside, spent walking and swimming, and constantly in the company of Alexandrov, he was inspired. In a 'dream' he envisaged the outlines of a comprehensive theory of topology and a solution to the dimension problem.

During the following academic year, 1921–1922, Urysohn proved many new theorems in point-set topology. He gave a course on topological continua and made it his practice to give students the proofs of new results immediately he discovered them himself. Simultaneously he announced his results in a series of notes to the Moscow Mathematical Society. By the spring of 1922, less than a year after he had begun, his dimension theory was settled, and in September Lebesgue presented the theory to the *Académie des Sciences* in Paris. When the full paper appeared in print several years later, it constituted our 'Landmark' in topology. Its contents are summarised in Table 1.

In approaching his task, Urysohn laid down several methodological principles. All his definitions had to be intrinsic. Moreover, he sought *local* definitions rather than global ones. This would allow the treatment of spaces which contained points of differing dimensions, like the 'disk with spike' defined as the subset $\{(x, y)\colon x^2 + y^2 = 1\} \cup \{(x, 0)\colon 1 \leqslant x \leqslant 2\}$ of E^2. Closed sets were his primary focus, but he noted that several theorems did not require this condition. He also saw that compactness, in the sense of Maurice Fréchet, defined in terms of infinite sequences and limit points (what is now called 'sequential compactness') was the pivotal assumption in most of his arguments [Pier, 1980]. Hence, it seemed natural to use the compact metric space as a base for his work rather than the more concrete Euclidean spaces. Indeed, Urysohn proclaimed the compact metric space to be the 'natural domain of existence' ('domaine naturel d'existence') for intrinsic topology, but, as Dale Johnson has remarked, Urysohn made the statement, not realising that even 'natural domains' change in the course of history and are altered to fit the current demands of mathematical research [Johnson, 1981, 229].

In seeking the definition of curve and surface expressed in the language of set theory, Urysohn set out a definition of an n-dimensional Cantorian manifold ('multiplicité Cantorienne') as an n-dimensional continuum, which remains connected after the removal of any closed subset of dimension $n - 2$. The construction effectively solved Egorov's problem of requiring intrinsic definitions of lines and surfaces since in Urysohn's terms, curves and surfaces are one- and two-dimensional Cantorian manifolds respectively. (The technical term 'Cantorian manifold' has been dropped in topology as 'manifold' is now reserved for a topological space that is locally Euclidean.) This definition was predicated on the intuitive meaning of 'dimension' itself.

Table 1. Summary by Chapters of Urysohn's paper. Part 2 starts at Chapter 3.

Chapter	**Page**	**Topics**
Introduction	30–64	Problems of topology, methodological principles. Summary of known results, concepts, terminology, and notation of topology.
1. Definitions	65–79	Definition of ε-separation (p. 65). Inductive definition of dimension *ind* (p. 66). Dimension is a topological invariant, and other consequences. Sets of dimension zero.
2. Preliminary study of dimension	79–137	Consequences of definition for subsets of Euclidean n-space. Particular results follow which presage general results (investigated in ch. 5): E^2, Cantorian lines; E^3, Cantorian surfaces. Definition of a Cantorian n-dimensional manifold (p. 124).
3. Examples	225–256	Exhibition of various test-bed examples of indecomposable continua.
4. Fundamental theorems	256–286	Concerning the dimension of closed sets: e.g. a closed set which is of dimension at least n cannot be decomposed into a countable (or finite) number of sets of dimension less than n (p. 260).
5. Euclidean spaces	286–316	Introduction to E^n. Relation between dimension and true order; $ind = dim$, that is, a necessary and sufficient condition for a closed set F to have $ind(F) = n$ is that its true order equals $n + 1$ (p. 301). Several important theorems: e.g., a closed domain of E^n is of dimension n, and the double boundary theorem (the set common to two domains in E^n has dimension $n - 1$ (p. 311)).
6. Decomposition of sets	316–351	The decomposition of sets into sets of zero dimension; Urysohn's inequality ((6) in text) (p. 317). Decomposition of n-dimensional sets into $n + 1$ sets of zero dimension. Theorems on closed sets and F_σ sets.
Supplementary notes	352–356	Condition for the intersection of sets to be of dimension n; a property of F_σ, G_δ sets.
	357–359	New notation, Table of contents.

Urysohn's notion of dimension is based on 'ε-separation'. A point x in a subset C of a compact metric space is ε-separated by a set B if there are mutually disjoint sets A, B and D such that i) $C = A \cup B \cup D$, ii) $x \in A$, iii) $A \cup B \subset S$ (an open ball centre x and radius ε), and iv) $(A \cap \overline{D}) \cup (\overline{A} \cap D) = \phi$. His definition of dimension is recursive: a point

x is of dimension n if it is *not* of dimension $< n$ with respect to C but can be ε-separated by a set B of dimension $< n$. That an isolated point is of dimension zero is obtained by defining the dimension of the empty set to be -1, the only set with this property. If no such n exists, the point x is said to be of infinite dimension. In modern terms Urysohn's definition is referred to as the small inductive dimension and is denoted by *ind*.

Urysohn made substantial progress in constructing a theory of *ind* based on Cantorian manifolds. A major step was his implicit establishment of the 'coincidence theorem'

$$ind\, X = Ind\, X = dim\, X \tag{1}$$

for compact metric spaces. An instance of his work is his proof of the 'addition theorem'

$$ind(Y \cup Z) \leqslant ind(Y) + ind(Z) + 1 \tag{2}$$

for separable spaces Y and Z, a theorem sometimes referred to as 'Urysohn's inequality' (p. 317). As a straightforward extension where X is the union of k subsets F_i, he found that

$$ind(X) \leqslant \sum_i ind(F_i) + (k-1). \tag{3}$$

In the special case that X is of dimension n and is decomposable into k subsets of dimension zero, it follows that $k \geqslant n+1$. He went on to show that such a decomposition of X may be achieved with exactly $n+1$ zero-dimensional sets.

Completing this work in the spring of 1923, Urysohn and Alexandrov set off on a journey through Europe, funding it by giving popular public evening lectures on the new relativity theory in several Moscow theatres. They visited Göttingen and met such leading lights as David Hilbert (1862–1943), Emmy Noether (1882–1935), and Richard Courant (1888–1972). There Urysohn became aware of Brouwer's dimension theory, and studying it, found a counter-example to one of Brouwer's results concerning the 'separation of points'. In September he lectured on this example to a meeting of the German Mathematical Union in Marburg.

In the following year, 1924, Urysohn and Alexandrov returned to Europe. They met with Hilbert again, who welcomed their joint paper on topological spaces for the *Mathematische Annalen*—and thanked them for a gift of caviar. In Bonn they met Hausdorff, yet another topologist who had started off in physics. They visited him in his home for mathematical seminars to discuss new ideas in topology, and alarmed him by their daily swimming of the Rhine. His *Grundzüge der Mengenlehre* of 1914 provided a basis for Urysohn's work, and he adopted Hausdorff's terminology, which he found systematic and complete. Continuing their tour, they journeyed to Holland to sit at the feet of Brouwer. Frequently a man who could dismiss ideas and their creators with withering scorn, Brouwer took the young Russian mathematicians into his confidence. Leaving Holland, the two 'P.S.'s' continued their European meanderings to France, to the coast of Brittany near Nantes. They stopped at Batz-sur-Mer (near La Baule) on the coast of Brittany, where on 17 August a calamity occurred. In the morning Urysohn began a new mathematics paper; but in the late afternoon, dismissive of the danger of swimming in rough seas, he was drowned.

4 KARL MENGER, MATHEMATICIAN

After a short while at the University of Vienna, Menger grew dissatisfied with the physics course and migrated to the Institute of Mathematics. During the 1920s the university possessed an enviable mathematical reputation, with a faculty including Wilhelm Wirtinger (1865–1945), Philipp Fürtwängler (1869–1940), and Hans Hahn (1879–1934).

In the spring of 1921, Hahn set up a seminar on problems associated with the theory of curves. Menger attended, and inspired by Hahn, quickly produced a paper titled 'New ideas concerning the concept of curve'. As Dale Johnson has noted, Menger's endeavour was only an informal description of curves embedded in Euclidean 3-space but it pointed the way to his future work [Johnson, 1981, 233–241]. With progress in prospect, Menger suffered a severe setback: he was afflicted with tuberculosis (Morbus Viennensis) and was forced to retire to a sanatorium. There he had time to reflect and to revise his notion of a curve.

Once recuperated, Menger returned to Vienna in April 1923, where he worked on his 'Landmark' paper. By the end of the year, he submitted it to the *Monatshefte für Mathematik und Physik*, a journal produced locally, of which Hahn was an editor. This first Part formed the basis of his thesis, and he gained his doctorate on 23 June 1924. A year later a second Part of the paper was submitted and published in 1926. Their contents are summarised in Table 2.

Menger approached his definition of dimension with a clear insight. Moreover, when approaching geometrical situations, with a view to shaping a theory, he confronted the theory in a physical way: 'We can think of curves as being represented by fine wires, surfaces produced from thin metal sheets, bodies as if they were made of wood. Then we see that in order to separate a point in the surface from points in a neighbourhood or from other surfaces, we have to cut the surfaces along continuous lines with a scissors' [Menger, 1925, 278; see Crilly and Moran, 2002, 146]. He could then translate his thoughts into a mathematical language. In this way, a curve for Menger was a connected subset K of a metric space with the property that each neighbourhood of its points contains a neighbourhood whose intersection with K is disconnected. In the first Part of his landmark paper he probed this definition and proved some theorems.

The fundamental definition at the base of Menger's dimension theory is, like Urysohn's, recursive. It is more immediate than Urysohn's, and though phrased in terms of sets, it is actually defined as a local concept: a subset M of a metric space is called n-dimensional if n is the smallest number with the property, that for each point m in M and to each neighbourhood $U(m)$ there exists a neighbourhood $U'(m) \subset U(m)$, so that the intersection of M with the boundary of $U'(m)$ is at most $(n-1)$-dimensional (pt. 1, 158). This illustrates why Menger took the dimension of the empty set to be -1, an important part of the definition. As with Urysohn, a subset that was not n-dimensional for any natural number n, was called infinite-dimensional.

While the first Part of Menger's paper consists of miscellaneous topics in the theory of curves, the second Part is almost entirely dedicated to dimension theory. In this he allowed the definition of dimension to extend beyond metric spaces and apply to topological spaces as defined by Hausdorff in terms of systems of neighbourhoods (pt. 2, 138).

Table 2. Summary of Menger's paper. The second Part begins at 'Introduction'.

	Pages	**Topics**
	148–149	Definition of a curve in a metric space.
	149–157	Properties of curves (e.g. if a compact curve K lying in a metric space is mapped by a continuous transformation to a metric space, its image is also a curve (p.149). If the union V of a countable number of curves is a compact continuum, then V is a curve (p. 152)).
	158	Recursive definition of dimension of a metric space.
	159–160	Properties of 0-dimensional sets. Curves which can be identified with 1-dimensional continua. A proof that the dimension of E^2 is 2. A geometric description of closed sets in E^2.
	160	Statement of independence from Urysohn.
Introduction	136–139	The dimension problem. Recursive definition of dimension for topological spaces which are capable of being defined by Hausdorff's axioms stated in terms of neighbourhood systems (p.138). Consequences of the definition of dimension; e.g. the dimension of a subset of an n-dimensional space is at most n-dimensional (p. 139).
Section 1	139–141	'Invariance of dimension' under homeomorphism (p. 140)).
Section 2	141–152	'Structure of n-dimensional sets'. Theorems: e.g. the set of all points of $\overline{M}$ in which M is at most (at least) k-dimensional, can be constructed from a $G_\delta(F_\sigma)$ set (p. 141).
Section 3	152–161	Properties of E^n and subsets. Theorems: e.g. The dimension of E^n (and each open subset of it) is n (p. 155); The complement of a compact closed subset of E^n which is at most $n-2$ dimensional, is connected (p. 156). Relationship between dimension and connectivity (pp. 157–161).

In a further Part of his paper with the same title, Menger [1929] sought to characterize the dimension function by a set of axioms. He gave five axioms, which a dimension function should satisfy, and showed that such a function defined on subsets of the plane was identical to *ind*. This axiomatic line of investigation was continued by Georg Nöbeling (b. 1907), Witold Hurewicz (1904–1956), Henry Wallman (1915–1992) and Ryszard Engelking.

5 THE JOINT CONTRIBUTION

Emerging from the aftermath of revolution and the first World War, Menger and Urysohn looked at the problem of dimension anew and independently constructed topological theories based on their definitions. Their problem was to build a theory around geometric sets of points involving definitions that could deal with the 'pathological' examples in topol-

ogy that had appeared since the time of Cantor. In the course of their work they built a solid foundation for the development of dimension theory, as an important part of point set topology, which by Urysohn and Menger's time was a semi-established subject broadly concerned with the properties of sets invariant under homeomorphism.

Neither Menger nor Urysohn arrived at their theory without guidance. As Dale Johnson has noted, the young mathematicians were guided by 'intellectual fathers' who had a knowledge of the problems and the attendant difficulties. In the case of Menger it was Hahn, and in Urysohn's case it was Egorov [Johnson, 1981, 240]. Both 'fathers' had connections with leading workers in the field of topology. A good friend of Brouwer, Hahn had worked on the theory of curves before the War, and would have been aware of his researches on dimension. Hahn himself had already shown that the essential property of a curve was its local connectedness, and this was a strong element in Menger's work. Egorov had posed the problem that set Urysohn off on his quest, but the indirect influence of Waclaw Sierpinski (1882–1969) on his development cannot be ignored. Technically an Austrian citizen he was stranded in Russia during the War but allowed to continue mathematical research in Moscow. He knew Egorov and collaborated with Luzin and produced papers on curves and continua. After the War, he returned to Poland, and worked on dimension theory independently of Menger and Urysohn.

Menger started on the problem without years of mathematical training, the area of point-set topology presenting problems that do not require a long pre-schooling. These pioneering problems clearly appeal to the 'tough-minded' specialist who requires a challenge, and this suited the young Menger keen to establish his name. Urysohn approached the problem with a rigorous mathematical training and the self-knowledge that he had gained his mathematical 'spurs' with a thesis on integral equations and an established research record. Menger enjoyed a long life and forged a high reputation, but, considering his life's brief span, Urysohn's accomplishment is breathtaking. Alexandrov spent the summers of 1925 and 1926 in Holland where he and Brouwer made Urysohn's work ready for posthumous publication. Of the Soviet school of mathematics, which grew so rapidly following the First World War, Alexandrov wrote of his lamented friend 'Pavel Samuilovich Urysohn was one of the greatest, if not the greatest, of these both in his talents and in his enthusiasm' [Arkhangelskii and Tikhomirov, 1998, 875].

6 THE IMPACT OF MENGER AND URYSOHN

The first impact of Urysohn's work was on Alexandrov himself. Knowing his work intimately Alexandrov went on to place some of his friend's results in a general setting. For instance, Urysohn had conjectured that a set which is the common boundary of *two* regions (a 'double boundary') in Euclidean n-space is an $(n-1)$-dimensional Cantorian manifold. For $n = 2$ it was known, and it was settled by Urysohn himself for $n = 3$. In 1927, Alexandrov proved it generally as a prologue to homological dimension theory.

In assessing their impact more generally, the pioneering aspect of their work should not be forgotten. For any who wanted to study dimension theory in the 1920s, there was only a scattered set of materials available. Brouwer's papers could be consulted, while Poincaré had published his sketchy thoughts in a philosophical journal. Lebesgue's 'covering dimension' was only fully investigated by its creator following the War. Thus Urysohn's

landmark paper fulfilled a need by offering extensive coverage of the embryonic theory. It was also written in a style that suggested further work. Its publication in the prestigious Polish journal, dedicated solely to this branch of topology, perhaps owes something to Luzin and Egorov's friendship with Sierpinski.

A continuation of Urysohn's landmark paper, covering a further 172 pages, was edited by Alexandrov and published in 1927. The total length of Urysohn's papers on Cantorian manifolds amounted to 417 printed pages. Menger's landmark papers were short by comparison, but he became prolific and his definition of dimension found its way into his papers on the theory of curves and surfaces. His book *Dimensionstheorie* [Menger, 1928] provided the first comprehensive treatment of the theory and made the theory generally available. It remained a standard reference even after the appearance of Hurewicz and Wallman's *Dimension theory* (1941) that gives a polished presentation of the theory in which the separable metric space provides the setting.

Following the appearance of the landmark papers, there were several significant departure points in dimension theory. During 1925–1927 the coincidence theorem, that all three classical definitions of dimension (*ind*, *Ind*, and *dim*) coincide for the class of separable metric spaces, was proved by Hurewicz and Lev Abramovich Tumarkin (1904–1974) independently. This theorem allowed topologists the flexibility to apply the most appropriate form of the dimension concept for these spaces, sure in the knowledge that these definitions were equivalent.

By the early 1930s dimension theory as applied to separable metric spaces was firmly established, but the tendency for mathematicians to generalise cast this work as 'classical'. Eduard Cech (1893–1960) modified the recursive definition of *ind* and defined dimension for topological spaces which included metric spaces as a special case. In a next step, the broadening of dimension theory beyond separable metric spaces, some of the fundamental identities were lost. An immediate casualty was the coincidence theorem, which fails for both compact spaces and for metric spaces generally [Engelking, 1968, 262–264]. For the broad classes of topological spaces with few restrictions there was not one unique value of topological dimension but potentially three different ones. These new features signalled a rich theory ahead.

At the Moscow International Conference on Topology in 1935, Alexandrov posed questions concerning the relationships between $ind\,X$, $Ind\,X$, $dim\,X$. For instance, what is the widest class of topological spaces for which $dim\,X \leqslant ind\,X$? In 1941, he proved this was true for compact spaces, but in 1949, A. Lunc and O.V. Lokucievskii each gave examples of compact spaces for which $dim\,X = 1$ and $ind\,X = 2$ [Alexandrov, 1955, 3]. Kiiti Morita (1915–1995) in 1950, and Yu.M. Smirnov (b. 1921) in 1951 proved the inequality for Lindelöf spaces and Morita for complete paracompact spaces, and it has been further extended. The reverse inequality, $dim\,X \geqslant ind\,X$, met with lesser success and the interest shifted to describe spaces for which equality holds. In 1952, Miroslav Katêtov (1918–1995) showed this was the case for metrizable spaces and by Morita two years later.

Urysohn's work inspired a generation of topologists in Russia. In 1951 Smirnov proved Urysohn's inequality (3)

$$ind(Y \cup Z) \leqslant ind(Y) + ind(Z) + 1, \tag{4}$$

where Y, Z are sets in a hereditary normal space X (and in 1963 it was established for normal spaces by Alexander Zarelua). In the West his work attracted topologists such as Edwin Hewitt (1920–1999) and C.H. Dowker (1912–1982). According to prominent Russian topologists, Urysohn's work really applied to the development after the 1950s, where his influence is pervasive [Arkhangelskii and Tikhomirov, 1998, 890].

Of all the definitions of topological dimension, *ind*, *Ind* and *dim* are the main ones, and Urysohn's and Menger's *ind* has met with the greatest success. 'None of the several other possible definitions of dimension has the immediate intuitive appeal of this one and none leads so elegantly to the existing theory' [Hurewicz and Wallman, 1948, 4]. The establishment of *ind* was far reaching, for it could be applied to subspaces of Hilbert space though results along these lines were still in their infancy around 1950.

We can now see that Menger and Urysohn's work may be regarded as complementary. Though equivalent, Urysohn's definition of dimension does not have the immediacy of Menger's, the form of the definition of local dimension that has achieved widespread use. In his compendious work, Urysohn gave a smooth presentation of a far greater number of fundamental results. Taken together, these two mathematicians provided a benchmark for dimension theory in the 1920s ready for further advance. When thinking of the birth of dimension theory, topologists correctly acknowledge the joint 'Urysohn–Menger' theory.

BIBLIOGRAPHY

Alexandrov, P.S. 1955. 'The present status of the theory of dimension', *American Mathematical Society translations*, *1*, 1–26.

Alexandrov, P.S., and Fedorchuk, V.V. with Zaitsev, V.I. 1978. 'The main aspects in the development of set-theoretical topology', *Russian mathematical surveys*, *33*, 1–53. [Signposts 1900–1978.]

Arkhangelskii, A.V. and Tikhomirov, V.M. 1998. 'Pavel Samuilovich Urysohn (1898–1924)', *Russian mathematical surveys*, *53*, 875–892.

Aull, C.E. and Lowen, R. (eds.) 1997. *Handbook of the history of general topology*, vol. 1, Dordrecht: Kluwer.

Cameron, D.F. 1982. 'The birth of Soviet topology', *Topology proceedings*, *7*, 329–378. [Biographical details on Urysohn not available elsewhere in English.]

Crilly, T. with Johnson, D. 1999. 'The emergence of topological dimension theory', in [James, 1999], 1–24.

Crilly, T. and Moran, A. 2002. 'Commentary on Menger's work on curve theory and topology', in [Schweizer et alii, 2002], 141–152.

Engelking, R. 1968. *Outline of general topology*, Amsterdam: North-Holland. [Trans. from Polish. Contains historical notes and references.]

Fedorchuk, V.V. 1998. 'The Urysohn identity and dimension of manifolds', *Russian mathematical surveys*, *53*, 937–974.

Hurewicz, W., and Wallman, H. 1941. *Dimension theory*, Princeton: Princeton University Press. [Rev. ed. 1948.]

James, I.M. (ed.) 1999. *History of topology*, Amsterdam: Elsevier.

Johnson, D.M. 1979, 1981. 'The problem of the invariance of dimension in the growth of modern topology, Part I' and 'Part 2', *Archive for history of exact sciences*, *20*, 97–188; *25*, 85–267.

Johnson, D.M. 2002. 'Commentary on Menger's work on dimension theory', in [Schweizer et alii, 2002], 23–32.

Katêtov, M. and Simon, P. 1997. 'Origins of dimension theory', in [Aull and Lowen, 1997], 113–134. [General exposition up to the 1920s.]

Koetsier, T. and van Mill, J. 1997. 'General topology, in particular dimension theory, in The Netherlands: the decisive influence of Brouwer's intuitionism', in [Aull and Lowen, 1997], 135–180.

Menger, K. 1925. 'Grundzüge einer Theorie der Kurven', *Mathematische Annalen, 95*, 277–306.

Menger, K. 1928. *Dimensionstheorie*, Leipzig and Berlin: Teubner.

Menger, K. 1929. 'Über die Dimension von Punktmengen III. Zur Begründung einer axiomatischen Theorie der Dimension', *Monatshefte für Mathematik und Physik, 36*, 193–218. [Repr. in [Schweizer et alii, 2002], 93–118.]

Menger, K. 1943. 'What is dimension?', *American mathematical monthly, 50*, 2–7.

Menger, K. 1979. *Selected papers in logic and foundations, didactics, economics*, Dordrecht: Reidel (Vienna Circle Collection, vol. 10).

Menger, K. 1994. *Reminiscences of the Vienna Circle and the Mathematical Colloquium* (ed. L. Golland and others), Dordrecht: Kluwer (Vienna Circle Collection, vol. 20).

Pier, J.-P. 1980. 'Historique de la notion de compacité', *Historia mathematica, 7*, 425–443.

Schweizer, B. et alii (eds.) 2002. *Karl Menger selecta mathematica*, vol. 1, Vienna: Springer. [Selection of papers and commentaries; does *not* include the landmark paper.]

Topology atlas. WWW source for obituaries, latest research news, etc.

CHAPTER 67

R.A. FISHER, *STATISTICAL METHODS FOR RESEARCH WORKERS*, FIRST EDITION (1925)

A.W.F. Edwards

This book is especially notable for its wide-ranging account of methods of statistical inference, and also for the wealth of applications made to biology.

First publication. Edinburgh and London: Oliver and Boyd, 1925 (Biological monographs and manuals (ed. F.A.E. Crew and D.Ward Cutler), vol. 5). ix + 239 pages + 6 pull-out tables. Print run: 1050 copies.

Later editions. 2nd 1928, 3rd 1930, 4th 1932, 5th 1934, 6th 1936, 7th 1938, 8th 1941, 9th 1944, 10th 1946, 11th 1950, 12th 1954, 13th 1958; all Oliver and Boyd. 14th (posthumous), in two variants: 1) Oliver and Boyd (ISBN 0-05-002170-2), 1970; 2) New York: Hafner; London: Collier–Macmillan, 1970.

Reprints. 10th, 13th and 14th editions, the last in *Statistical methods, experimental design and scientific inference*, Oxford: Oxford University Press, 1990 (ISBN 0-19-852229-0).

French translation. Méthodes statistiques adaptées à la recherche scientifique, Paris: Presses Universitaires de France, 1947.

German translation of the 12th ed. Statistische Methoden für die Wissenschaft, Edinburgh: Oliver and Boyd, 1956.

Italian translation. Metodi statistici ad uso dei ricercatori, Turin: Unione Tipografico, 1948.

Spanish translation of the 10th ed. Metodos estadisticos para investigadores, Madrid: Aguilar, 1949.

Japanese translation. Kenkyuusyano tameno toukeiteki houhou, Tokyo: Morikita, 1952. [Repr. 1970.]

Russian translation of the 12th ed. Statisticheskie metody dlya issledovatelei, Moscow: Gosstatizdat, 1958.

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

Related articles: Bayes (§15), Laplace on probability (§24), Pearson (§56), Shewhart (§72).

1 THE AUTHOR

Although R.A. Fisher (1890–1962) was a first-rate mathematician, *Statistical methods for research workers* contains no advanced mathematics. Reviewers of the first edition objected to its narrow focus, its lack of proofs, its reliance on genetical examples, and its biological emphasis generally [Box, 1978]. Its greatness stems not from its mathematics but from the revolutionary novelty of its approach to statistical inference combined with its wealth of practical advice on the actual analysis of data by the new methods. Rarely has a book contained so much of value to the beginner and the advanced worker at the same time. Fisher's own description of the book, ten years after its publication, was 'a connected account of the applications in laboratory work of some of the more recent advances in statistical theory' [Fisher, 1935].

Fisher was taught 'all the pure, mathematics I know' by G.H. Hardy, whose textbook *A course of pure mathematics* (1908) had just been published when the young student came up to Cambridge to read for the Mathematical Tripos, in which he was placed in the first class in 1912. His contributions to mathematics itself lie in the fields of statistical distribution theory, combinatorial theory (especially the enumeration of Latin squares) and design theory generally. He initiated stochastic diffusion theory in 1922, and when A.N. Kolmogorov referred to 'das wundervolle Buch von R.A. Fisher' it was not to *Statistical methods* but to *The genetical theory of natural selection* [Fisher, 1930a], in which Fisher had employed it [Kendall, 1990]. In 1934 Fisher published the idea of a randomized or 'mixed' strategy in the theory of games, independently of von Neumann. But for most of his working life he was a professor of genetics, first in London and then at Cambridge, and in evolutionary theory he is acknowledged as 'the greatest of Darwin's successors' ([Dawkins, 1986]; see [Edwards, 1990]).

Ronald Aylmer Fisher was born in London on 17 February 1890, the son of George Fisher, a fine-art auctioneer, and his wife Katie [Box, 1978]. His twin brother was stillborn. At Harrow School, which he entered in 1904 as a scholar, he distinguished himself in mathematics despite being handicapped by poor eyesight that prevented him working by artificial light. His teachers used to instruct him by ear, and Fisher developed a remarkable capacity for pursuing complex mathematical arguments in his head. This manifested itself later in life in an ability to reach a conclusion whilst forgetting the argument, to handle complex geometrical trains of thought, and to develop and report essentially mathematical arguments in English (only for students to have to reconstruct the mathematics later). Fisher's early interest in natural history was reflected in the books chosen for special school prizes at Harrow, culminating in his last year in the choice of the complete works of Charles Darwin in 13 volumes.

Fisher entered Gonville and Caius College, Cambridge, as a scholar in 1909. After graduating in 1912 he spent a postgraduate year in the Cavendish Laboratory, Cambridge, studying the theory of errors under F.J.M. Stratton and statistical mechanics and quantum theory under J.H. Jeans (later Sir James Jeans).

Prevented from entering war service in 1914 by poor eyesight, Fisher taught physics and mathematics in schools for the duration of the war, and in 1919 he was appointed Statistician to Rothamsted Experimental Station, an agricultural research station at Harpenden, north of London. From 1920 to 1926 he was also a non-resident Fellow of Gonville and Caius College, to which he gave pride of place on the title page of *Statistical Methods*. In 1933 he was elected to succeed Karl Pearson (1857–1936) at University College London as Galton Professor of Eugenics (that is, of Human Genetics, as it later became), and in 1943 he was elected Arthur Balfour Professor of Genetics at Cambridge and once more a Fellow of Gonville and Caius College.

After he retired in 1957 Fisher traveled widely, spending his last few years in Adelaide, Australia, as an honorary research fellow of the C.S.I.R.O. Division of Mathematical Statistics. He died there of a post-operative embolism on 29 July 1962. His ashes lie under a plaque in a side aisle of Adelaide Cathedral.

Fisher married Ruth Eileen Guinness in 1917 and they had two sons and six daughters, and a baby girl who died young. He was elected a Fellow of the Royal Society of London in 1929 (as a mathematician) and was created Knight Bachelor by Queen Elizabeth II in 1952 for services to science. He was the founding President of the Biometric Society (now the International Biometric Society) in 1947, and served as President of the Royal Statistical Society (of the U.K.), of the Genetical Society of Great Britain, and of his Cambridge college, Gonville and Caius. He received many honorary degrees and accepted the honorary membership of many academies at home and abroad, and was awarded all the principal medals of the Royal Society, the Royal (1938), the Darwin (1948) and the Copley (1956).

Fisher's papers number nearly three hundred, and he also wrote many reviews, particularly in *The eugenics review* between 1915 and 1935, and letters to journals. In genetics, his book *The genetical theory of natural selection* [Fisher, 1930a] was followed by *The Theory of inbreeding* [Fisher, 1949]. Several statistics books were offshoots of *Statistical methods*, and will be mentioned below. He was an accomplished formal lecturer as may be seen from his many presidential and similar addresses, and an occasional broadcaster on scientific topics. The same cannot be said of his lectures to students, which required intense concentration and subsequent interpretation.

Small of stature, with thick glasses and a beard (Figure 1), Fisher did not suffer fools gladly. He was a skilled controversialist in conversation, but his quick temper sometimes rendered further discussion impossible. His contemporaries divided cleanly into those who regarded him with awe and affection and gratitude for the generosity with which he offered ideas to them, and those who found him tetchy, difficult and remote. Especially as a professor at Cambridge he showed great interest in the few students who passed through his small department, and he always enjoyed the company of young people.

Of the many thumb-nail sketches which his greatness has inspired, perhaps the following comment on Fisher by the Cambridge cosmologist Sir Fred Hoyle [1999] contains the closest likeness in the smallest span:

> I am genuinely sorry for scientists of the younger generation who never knew Fisher personally. So long as you avoided a handful of subjects like inverse probability that would turn Fisher in the briefest possible moment from ex-

Figure 1. Fisher in 1924, at the time of writing the first edition of his book. Photograph courtesy of Joan Fisher Box.

> treme urbanity into a boiling cauldron of wrath, you got by with little worse than a thick head from the port which he, like the Cambridge mathematician J.E. Littlewood, loved to drink in the evening. And on the credit side you gained a cherished memory of English spoken in a Shakespearean style and delivered in the manner of a Spanish grandee.

2 WRITING *STATISTICAL METHODS*

At Cambridge Fisher's introduction to statistical procedures was through the astronomer F.J.M. Stratton, one of his teachers in Gonville and Caius College. Not only did Stratton give an undergraduate course of lectures on 'Combination of observations' which Fisher almost certainly attended, but Fisher studied under him and Jeans at the Cavendish Laboratory for a postgraduate year 1912–1913. Although no notes for the lectures are known,

(Sir) David Brunt wrote a book *The combination of observations* in whose preface he wrote 'I have to acknowledge my indebtedness to Mr F.J.M. Stratton, of Gonville and Caius College, Cambridge, to whose University lectures I owe most of my knowledge of the subjects discussed in this book, and upon whose notes I have drawn freely' [Brunt, 1917]. The book refers to very little beyond Karl Pearson's work (§56), but it does notice the second edition of G. Udny Yule's *An introduction to the theory of statistics* (1912). These two books, with their heavy emphasis on the normal distribution, the method of least squares, correlation and (in Yule's case) contingency, may be assumed fairly to reflect Fisher's undergraduate knowledge (Whittaker and Robinson's *The calculus of observations* did not appear until 1924).

But Fisher was soon outpacing his teachers. Even as an undergraduate, encouraged by Stratton, he had in 1912 published a paper foreshadowing his later introduction of the method of maximum likelihood. This led him to an interest in the z-distribution (essentially the modern t-distribution) described, but not formally derived, in the path-breaking paper by 'Student' (W.S. Gosset) in 1908 that launched the statistical theory of small samples. Fisher was soon able to derive the distribution using n-dimensional geometrical reasoning.

By 1923 Fisher was already beginning to achieve the leading position in statistics which he was to hold at least until the outbreak of the Second World War in 1939. His 1922 *Philosophical transactions* paper 'On the mathematical foundations of theoretical statistics' was only one of the key papers in a period of five years marked by his introduction into statistics of many of the words and phrases which were to dominate the field: *variance*, *analysis of variance*, *degrees of freedom*, *efficiency*, *sufficiency*, *consistency*, *likelihood*, *method of maximum likelihood*, *location and scale* and *statistic* [Fisher, 1922].

Fisher started to write *Statistical methods* in the summer of 1923, and it was almost complete by the middle of 1924. The preface to the first edition is dated February 1925, and Cambridge University Library received its copy on 1 July. Fisher was in Canada from the end of July 1924 to the beginning of September and asked Gosset to read the proofs. One consequence of this was Gosset's suggestion on returning the proofs on 20 October that the all-important statistical tables could be folded 'into the book but when in use they could be folded out', and for the first six editions the tables in the text were duplicated in this way at the back of the book [Gosset, 1970].

The book was originally to be called *Statistics for biological research workers*, as may be seen in the first list of the books in the series 'Biological Monographs and Manuals' in which it appeared. This series, from the Edinburgh publishers Oliver and Boyd, was edited by F.A.E. Crew of Edinburgh and D. Ward Cutler of Rothamsted; but there was more than just Rothamsted to connect them to Fisher, for all three men were leading participants in the affairs of the Eugenics Society. In October 1924 Fisher was one of the Honorary Secretaries and both Crew and Cutler members of the Council. Crew became Professor of Animal Genetics in the University of Edinburgh and lived until 1973, whilst Ward Cutler died young in 1941 whilst still head of the Microbiology Department at Rothamsted.

Robert Grant was one of the partners in Oliver and Boyd and corresponded with Fisher in 1950 about the proposal by the editor of the *Journal of the American Statistical Association* to mark the silver jubilee of *Statistical methods* with 'one or two articles on the character and consequences of that volume' [Bennett, 1990]. He reminisced: 'It all takes my mind back to that day when Frank Crew called relative to your manuscript, how he

spoke of its quality, the formative work that it contained, and urged publication if only on the grounds that statistics in future would and must form part of research work in every science'. Fisher replied: 'It was Cutler who approached me, probably after consulting Crew, and certainly he came at the right moment, for I did not have to do any mathematical research *ad hoc*, but only had to select and work out in expository detail the examples of the different methods proposed'.

Ironically the first book of the series was by Lancelot Hogben [1924], for in later years Hogben was to be a leading critic of the entire corpus of Fisherian statistical inference. His magnum opus *Statistical theory*, though largely ignored by the statistical world, was a refreshing reminder that there is still much to be debated about inference. Of its many tilts at Fisher, this one was specific to *Statistical methods* [Hogben, 1957]:

> In *Statistical Methods for Research Workers*, destined to be the parent of a large fraternity of manuals setting forth the same techniques with exemplary material for the benefit of readers willing—and, as it transpired, only too anxious—to take them on trust, Fisher's formulation of the rationale of the significance test neither discloses a new outlook explicitly not clarifies views expressed by his predecessors. All that is novel is a refinement of the algebraic theory of the sampling distributions—with one notable exception embraced by Pearson's (1895) system of moment-fitting curves.

The language is eerily reminiscent of some of Fisher's own outbursts in later life, and effectively disguises such elements of truth as it contains.

Of the ten titles in 'Biological monographs and manuals' only *Statistical methods* survived into a second edition, but to the end of its life it was still being described as one of the series. Not until the 13th edition (1958) was the inclusion of a list of the other books in the series discontinued.

3 CONTENT OF THE FIRST EDITION

The first edition of *Statistical methods* is a handsomely-produced volume, 6×9 inches, bound in dark blue cloth with 'Biological Monographs and Manuals' printed in black on the front. Printing was undertaken in the publisher's own works. The lines of type are set at 15 point making for an 'open' appearance to each page. Fisher himself suffered from poor eyesight all his life, and remarked, in connection with *The genetical theory of natural selection*, 'Fairly large print is a real antidote to stiff reading' [Bennett, 1983]. The print run was 1050 copies (this figure and others quoted for subsequent editions are taken from [Yates, 1951]). The contents are summarised in Table 1.

The book opens with a long introductory chapter surveying the field of statistics from Fisher's new viewpoint. (The bold type is Fisher's, in the style of both [Yule, 1912] and [Brunt, 1917].) Section **1: The Scope of Statistics**—'Statistics may be regarded as (i.) the study of **populations**, (ii.) as the study of **variation**, (iii.) as the study of methods of the **reduction of data**'. The section ends: 'It is the object of the statistical processes employed in the reduction of data to exclude [the] irrelevant information, and to isolate the whole of the relevant information contained in the data'. Much of this opening material is taken directly from [Fisher, 1922] 'On the mathematical foundations of theoretical statistics'.

Table 1. Contents by chapters of Fisher's book, first edition (1925).

Ch.	**Pages**	**Contents**
I	26	Introductory.
II	16	Diagrams.
III	34	Distributions.
IV	24	Tests of goodness of fit, independence and homogeneity; table of χ^2.
V	37	Tests of significance of *means*, differences of means, and regression coefficients.
VI	36	The correlation coefficient.
VII	35	Intraclass correlations and the analysis of variance.
VIII	22	Further applications of the analysis of variance.

In Section **2: General Method, Calculation of Statistics** Fisher introduces his vital distinction between the **parameters** of the population distribution which are to be estimated and the **statistics** calculated from the data which are to be used for the purpose. Here we meet the singular word 'statistic' which he had coined in 1921:

> The problems which arise in the reduction of data may thus conveniently be divided into three types: (i.) Problems of **Specification**, which arise in the choice of the mathematical form of the population. (ii.) Problems of **Estimation**, which involve the choice of method of calculating, from our sample, statistics fit to estimate the unknown parameters of the population. (iii.) Problems of **Distribution**, which include the mathematical deduction of the exact nature of the distribution in random samples of our estimates of the parameters [. . .].

Here in these introductory pages we already find the great clarification which Fisher had brought to statistical inference in the preceding few years, especially by distinguishing clearly between a *parameter* and its *estimate*.

Next Fisher states that he believes the method of **Inverse Probability** 'is founded upon an error, and must be wholly rejected'. (It was not until 1950 that the word 'Bayesian' was coined, by Fisher himself, to refer to inverse probability (compare §15.5). It has now completely replaced the earlier phrase; see [David and Edwards, 2001].) Fisher's methods would rely on quite different probability arguments, and by making entirely clear his rejection of Bayesian methods he was sweeping away a confusion which had permeated statistical thinking from C.F. Gauss and P.S. Laplace right up to Pearson and Francis Edgeworth at the end of the 19th century. All these authors had intermingled Bayesian and non-Bayesian arguments, but Fisher's ambition and intention was to construct a strictly non-Bayesian methodology for the 'reduction of data'.

But 'This is not to say that we cannot draw, from knowledge of a sample, inferences respecting the population from which the sample was drawn, but that the mathematical concept of probability is inadequate to express our mental confidence or diffidence in making such inferences, and that the mathematical quantity which appears to be appropriate for measuring our order of preference among different populations does not in fact obey the

laws of probability. To distinguish it from probability, I have used the term **"Likelihood"** to designate this quantity'. Notwithstanding this bold introduction of a concept but four years old, likelihood itself does not feature strongly in the book other than in the method of maximum likelihood, but in 21st-century statistical theory it is one of Fisher's principal legacies [Edwards, 1972].

Having introduced the idea that a statistic, computed from the data, is to be used to estimate a parameter, it is necessary to learn how to chose between competing statistics. Section **3: The Qualifications of Satisfactory Statistics** addresses this problem. A statistic must be **consistent**, that is, converge on the true value of the parameter when the sample is large, and it should be as **efficient** as possible. It should use all the 'relevant information' available in the observations. Fisher had first introduced these words and phrases in 1922, and although in *Statistical methods* 'information' is not explicitly defined in the first edition, it is implicitly taken to be inversely proportional to the sampling **variance** of the estimate. In a contemporary paper 'Theory of statistical estimation' Fisher [1925] defined 'information' explicitly, thus giving the word a technical meaning three years before R.V.L. Hartley used it in its 'Shannon' sense. Fisher information, as it is often now called, is a measure of informativeness about something specific, the value of a parameter, whilst Shannon information is a measure of the capacity of a channel to transmit a message whether informative or not. Shannon's refers to the medium, Fisher's to the message. Sometimes it is possible to find a statistic which is not only efficient but which can also be shown to use all the information even in a small sample; it is then said to be **sufficient**, another 1922 coining. Such statistics, where they exist, Fisher asserted could be found by his **Method of Maximum Likelihood** (1922 again) which, in other cases, would at least uncover an efficient statistic.

Thus does Fisher, in a few sentences, set out his project. To readers of Yule and Brunt it will have appeared a great mystery. To Fisher's biological colleagues, unversed in least squares, finite differences, and other concepts inherited from astronomy and geodesy, it was a revelation.

Section **4: Scope of this Book** describes in some detail how the book is constructed. After Fisher's remark 'The book has been arranged so that the student may make acquaintance with these three main distributions in a logical order [χ^2, t and z], and proceeding from more simple to more complex cases', the exasperated anonymous original owner of my copy of the first edition has written 'and with the minimum of logical explanation'. Further on he writes in the margin 'Proof? Why not even refer to the paper which contains a proof?'. It was to be a common complaint from mathematicians. The statistician G.A. Barnard first met Fisher in 1933 when in his last year at school, and remarked that he had been looking for mathematical texts on statistics without success. Fisher pointed to a copy of *Statistical Methods* and said: 'I believe you are a mathematician. You'll find in this book a lot of statements given without proof. If you're a mathematician you should be able to prove these things for yourself. If you work through the book doing that, you'll learn mathematical statistics'. Barnard did [Barnard, 1990].

Section **5: Mathematical Tables** explains the sources of the tables included in *Statistical Methods*. Fisher made a fundamental change in the format of statistical tables by tabulating the value of the variate for a range of values of the probability rather than the other way round. Thus Table III for Pearson's χ^2 tabulates it for values of the probability

P (.99, .98, .95, . . . , .10, .05, .02, .01) and the degrees of freedom n (from 1 to 30) instead of tabulating P for different values of n and χ^2 as had Karl Pearson. This, coupled with the use of the novel phrases 'test of significance' (p. 43), 'significance level' (p. 157) and 'percentage point' (p. 198), drove the entire statistical community away from quoting P at the conclusion of an analysis towards asserting whether or not the result was 'significant' at a certain 'level', such as 1% or 5%. In Chapter III Fisher goes so far as to say 'It is convenient to take this [5%] point as a limit in judging whether a deviation is to be considered significant or not'. Hald [1998] famously described Fisher as 'the statistical magistrate of our time'. A further, and profound, consequence of this new format for statistical tables was the ease with which it enabled confidence intervals to be computed, which Egon Pearson [1990] thought might have been instrumental in the emergence of the confidence theory. In this he was echoing a thought of Fisher's in his Harvard Tercentenary Lecture 'Uncertain inference' [Fisher, 1936]. He also confirms the story that part of the reason for the new format might have been his father's reluctance to allow Fisher to reproduce the *Biometrika* copyright tables for fear of damaging their sales.

Section **6** has no title, and sits oddly in this first chapter, as if an afterthought. It consists of an example of the application of the method of maximum likelihood to the estimation of a genetical recombination fraction. As we shall see, Fisher later expanded it and placed it elsewhere.

Seven further chapters follow this, practising what the 'Introductory' Chapter I has preached. Chapter II covers 'Diagrams'—the very first of which charts the growth of baby Harry, the Fishers' second child, born in May 1923. In Chapter III, 'Distributions', the Normal, Poisson and Binomial distributions are introduced, with the minimum of mathematics but with examples of fitting them to data. A typical example of the author's efforts to explain mathematical concepts in words for the benefit of his biological readers is his description of the Normal distribution as having 'frequencies given by a definite mathematical law, namely, that the logarithm of the frequency at any distance x from the centre of the distribution is less than the logarithm of the frequency at the centre by a quantity proportional to x^2'. The formula is then revealed, one senses rather reluctantly.

Chapter IV introduces 'Tests of goodness of fit, Independence and Homogeneity' based on the χ^2 distribution. As usual, the instruction is by means of examples and no formula is given for the χ^2 distribution function. In Chapter V we move on to 'Tests of Significance of *Means*, Differences of Means, and Regression Coefficients'. Tests based on the Normal distribution soon give way to small-sample tests based on 'Student's' t-distribution. Characteristically, Fisher does not bother to refer to the fact that he was the mathematician who first rigorously derived this distribution. In the fourth edition (1932) Fisher added a '**Historical Note**' to Chapter I in which he said " 'Student's" work was not quickly appreciated, and from the first edition it has been one of the chief purposes of this book to make better known the effect of his researches, and of mathematical work consequent upon them'. An interesting occurrence of the word 'confidence' given its later use in 'confidence interval' occurs in Chapter V: 'In this case we can not only assert a significant difference, but place its value with some confidence at between 4 and 5 inches'.

'The Correlation Coefficient' is introduced in Chapter VI by means of the data of Karl Pearson and Alice Lee on the stature of 1376 father–daughter pairs. The formula for the correlated bivariate normal surface is given and the method of calculation of the correlation

coefficient demonstrated. The partial correlation coefficienti is described. The significance test for an observed correlation, involving 'Student's' distribution, is given, and Fisher shows by an example how misleading it would be to assume a Normal distribution of error for the correlation coefficient. To handle the sampling distribution of the coefficient when the population is itself correlated Fisher introduces his z-transformation, once again omitting to mention that it was he who derived the exact sampling distribution in 1915 that led him to suggest this transformation.

In Chapter VII 'the analysis of variance' makes its first appearance, via a discussion of 'Intraclass Correlations'. Here one may see the transition between the Pearsonian emphasis on correlation and the Fisherian emphasis on the analysis of variance taking place before one's very eyes. Fisher had only invented the terms 'variance' and 'analysis of variance' seven years earlier, in 1918 [David and Edwards, 2001]. Chapter VIII 'Further Applications of the Analysis of Variance' takes the story a step further, first into regression and then into the all-important area of agricultural experimentation. Section **48. Technique of Plot Experimentation** starts 'The statistical procedure of the analysis of variance is essential to an understanding of the principles underlying modern methods of arranging field experiments'. Randomisation is introduced as the prerequisite for the validity of the subsequent tests of significance. It was Fisher's first publication of this point of view. (In the third edition Fisher added to the Preface the remark: 'As has sometimes occurred before with the inclusion of new results, reference to the demonstration cannot yet be given, since no demonstration has yet been published'.) A 5×5 Latin Square makes its appearance. And there the book suddenly stops, just as it touches on these path-breaking developments at Rothamsted which were to lead to a revolution in agricultural experimentation and indeed to experimentation in science generally.

4 SUBSEQUENT EDITIONS AND BOOKS

Fisher produced a new edition of *Statistical methods* roughly every two years throughout its life, and in each Preface he indicated the sections that he had added or altered. The full collection is shown in Figure 2. Prefaces of the later editions may therefore be consulted for the details, and here we only note major changes.

In the second edition (1928, 1250 copies) Fisher removed Section **6** of the 'Introductory' Chapter I and expanded it into a new Chapter IX 'The Principles of Statistical Estimation', partly at the suggestion of Gosset. Generations of statistical geneticists have learnt their estimation theory from it. The same material appeared in a contemporary paper with B. Balmukand [1928]. We find statistical 'information' now clearly defined.

In the third edition (1930, 1500 copies) Fisher added a separate bibliography of his own statistical publications. This subsequently grew year by year, for his output never slackened.

The fourth edition (1932, 1500 copies) contained a long addition to Chapter VIII, a Section **49.1. The Analysis of Covariance**. The very word 'covariance' had only been coined two years earlier (by Fisher himself, in *The genetical theory of natural selection*). Of particular interest to students of the logic of statistical inference is the effect on *Statistical methods* of Fisher's discovery and advocacy of *fiducial probability* in [1930b]. In Chapter I

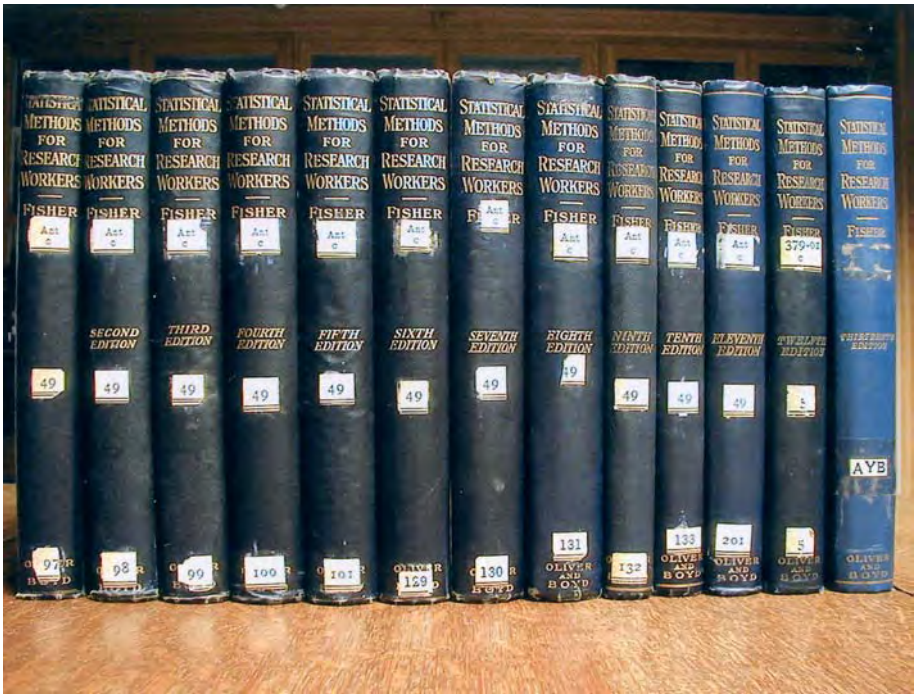


Figure 2. A population of 13 varying editions (Cambridge University).

he could no longer say that ‘the mathematical concept of probability is inadequate’ (for the full quotation, see section 3 above) without adding ‘in most cases’ and explaining that the interpretation of probabilities established by tests of significance ‘as probability statements respecting populations’ constituted ‘an application unknown to the classical writers on probability’.

Also for this edition Fisher removed Section **5. Mathematical Tables** from Chapter I and replaced it by a **Historical Note** intended to be ‘of value to students who wish to see the modern work in its historical setting’ and correcting misapprehensions ‘ascribing to the originality of the author methods well known to some previous writers, or ascribing to his predecessors modern developments of which they were quite unaware’. Fisher has sometimes been criticized for the inaccuracy of his historical writing, but he never withheld credit where credit was due, in this section most notably when mentioning C.F. Gauss: ‘He perceived the aptness [for estimation] of the Method of Maximum Likelihood, although he attempted to derive and justify this method from the principle of inverse probability. The method has been attacked on this ground, but it has no real connection with inverse probability’.

The print run for the fifth edition (1934) was 1500 copies. It contained new material on the analysis of the 2×2 contingency table.

In 1935 Fisher’s second statistical book *The design of experiments* appeared. As he said in the preface: ‘In 1925 the author wrote a book (*Statistical methods for research*

workers) with the object of supplying practical experimenters and, incidentally, teachers of mathematical statistics, with a connected account of the applications in laboratory work of some of the more recent advances in statistical theory' and he went on to explain how the new work had grown out of the eighth chapter of the old. Henceforth *Statistical methods* no longer needed additions on this subject, and the sixth edition (1936, 2000 copies) could refer to the new book, though Chapter VIII was retained. One notable coining in *The design of experiments* was the phrase 'null hypothesis', missing from *Statistical methods*.

The sixth edition, like *The design of experiments*, also made free use of the word 'fiducial'. To the fourth-edition statement quoted above Fisher added: 'To distinguish such statements as to the probability of causes from the earlier attempts now discarded [that is, inverse probability], they are known as statements of **Fiducial Probability**', and instead of the mean of a normal distribution being 'likely' to lie within the limits defined by the sample mean $\pm$ two standard errors it becomes 'probable, in the fiducial sense'.

For the seventh edition, in 1938 (2000 copies), the duplicate fold-out copies of the statistical tables were dropped from the end of the book, made redundant by the publication in that year of another *Statistical methods* off-shoot: *Statistical tables for biological, agricultural and medical research*, produced jointly with Fisher's successor at Rothamsted, Frank Yates. In this edition the size had increased by more than 50% over the first edition.

The eighth edition appeared during the war, in 1941 (2250 copies), and perhaps for this reason was reset with closer lines (from 15 to $13\frac{1}{2}$ point, a 10% increase in capacity for the written material). Section numbers appeared in the running heads of each page for the first time (as had been suggested by Gosset in 1924), and Ward Cutler, who had died, ceased to be listed as joint editor with Crew; but little else changed.

Subsequent editions showed rather modest changes or none at all. The original first-edition preface was dropped for the ninth (1944, 2000 copies). The first post-war edition, the tenth, sold well (1946, 3000 copies; reprinted 1948, 1500 copies; my copy of the reprint, though claiming to be published by Oliver and Boyd, has 'Hafner Publishing Company Inc., New York' on the title page). The 11th edition in 1950 had a print run of 7500 copies. The 12th was published in 1954 and the 13th (and last in Fisher's lifetime) in 1958. No figures for their print runs have been published, but the 13th was reprinted in 1963 and 1967.

In the preface to the 13th edition Fisher added a penultimate paragraph occasioned by the publication in 1956 of the third of his books which can be described as offshoots from *Statistical methods*, 'one devoted to the logic of induction, under the title *Statistical Methods and Scientific Inference*' (referring to it as '1957'; Fisher rather often made mistakes of one year in referring to his own books and papers). The paragraph is notable for its Parthian shot at the 'flood of literature [which since the middle of this century] has appeared bearing on statistical methods'. 'The authors are largely in mathematical teaching departments, and better trained as mathematicians than some of their predecessors. Too often, however, their experience has not included the training and mental discipline of the natural sciences, and much space is given to the trivial and the irrelevant'. It was a familiar refrain.

The posthumous 14th edition appeared in two versions, both dated 1970. The 'British' version was an Oliver and Boyd paperback. It opens with a 'Preface to the Fourteenth Edition' carrying the name of E.A. Cornish of Adelaide, who prepared it 'from notes left by Sir Ronald Fisher', though in fact the preface is simply Fisher's from the 13th edition with

an inserted paragraph mentioning the additions. The 'American' version was published by Hafner in New York. It reprints the preface to the 13th edition, unsigned, with a separate note conveying the information about the new material. The type-setting of the first batch of additional material in the two versions differs, the British version missing some parentheses in a formula on page 165, and the American having a word italicized. That of the second batch, Section **57.4**, appears identical. The natural inference is that the American version has been produced from the British with corrections. Both are copyright 'University of Adelaide', Fisher's literary executor.

The 1990 single-volume reprint *Statistical methods, experimental design and scientific inference* of all three of Fisher's statistical books uses the 'American' edition but with a further variant of the page 165 formula, though the note about the new material is now signed by J.H. Bennett, of Adelaide, who also expanded the Contents so as to record all the section headings. This edition records a 1973 reprint of the 14th edition. There is a foreword to the whole volume by Yates.

5 IMPACT OF THE BOOK

That Fisher is the father of modern statistics no-one will dispute, but to gauge the contribution of *Statistical methods* itself requires the thought-experiment of imagining it not to have been written. This would have made little difference to the dissemination of Fisher's more theoretical work, which would have still formed the backbone of 20th-century mathematical statistics; but the effect of that work on applied statistical practice would have been felt more slowly, particularly in biology, including genetics, medicine and the design of agricultural experiments. If, as seems only just, one includes *The design of experiments* in the assessment, the impact of *Statistical methods* throughout the biological sciences was profound and permanent, and from biology the influence spread rapidly into the social sciences. To a greater extent than Fisher ever wished, the test of significance became the *sine qua non* of received practice, demanded by the editors of journals even when estimation would have been more appropriate.

The book had as great an impact on teachers as it had on experimentalists, as may be seen by comparing the texts of Yule, Brunt (both mentioned above) and A.L. Bowley [1926] with those that followed *Statistical methods*, especially G.W. Snedecor's *Statistical methods applied to experiments in agriculture and biology* [1937]. Fisher spent the summer of 1931 at Iowa State University, Ames, at the invitation of Snedecor, giving three lectures each week based on *Statistical methods*. Similarly in 1936 he lectured on design from *The design of experiments*, which had been published the preceding year. The influence of these lectures at the leading American statistical laboratory, and the further contacts which resulted, was immense. So too was the influence of H. Hotelling at Columbia University, who reviewed no fewer than seven editions of *Statistical methods*, writing at the outset 'The author's work is of revolutionary importance and should be far better known in this country' [David, 1998].

In India too *Statistical methods* was particularly influential. Mahalanobis [1938] writes: 'This book has probably done more than anything else to make research workers in most diverse fields of study familiar with the practical applications of modern statistical methods, and to create a statistical attitude of mind among the younger generation of scientists'.

Of all the eulogies of this great book none is more extensive and better informed than that of [Yates, 1951]. on the occasion of its silver jubilee. It starts:

> It is now twenty-five years since R.A. Fisher's Statistical Methods for Research Workers was first published. These twenty-five years have seen a complete revolution in the statistical methods employed in scientific research, a revolution which can be directly attributed to the ideas contained in this book, and which has spread in ever-widening circles until there is no field of statistics in which the influence of Fisherian ideas is not profoundly felt.

6 EPILOGUE

It may not be inappropriate to end with a personal anecdote. I was one of only two students studying genetics in Fisher's Cambridge department during his last year as Professor (1956–1957). I had previously attended an introductory course in statistics (by H.E. Daniels) and soon found that Fisher's lectures, especially on linkage, demanded a much deeper knowledge. At the end of one lecture I asked his advice and he said, rather quizzically, that he had written 'one or two books on statistics' and perhaps I might like to consult them. I bought all three (and happily had him autograph them) and devoured the 12th-edition *Statistical methods* with unbridled enthusiasm. One of the effects of reading the book again 46 years later has been to remind myself of just how much my own statistical education came from this one short book, which itself, though already then 31 years old, was such an inspiration. Many others can tell similar stories.

BIBLIOGRAPHY

Barnard, G.A. 1990. 'Fisher: A retrospective', *Chance, 3*, 22–28.

Bennett, J.H. (ed.) 1983. *Natural selection, heredity, and eugenics*, Oxford: Clarendon Press.

Bennett, J.H. (ed.) 1990. *Statistical inference and analysis: Selected correspondence of R.A. Fisher*, Oxford: Clarendon Press.

Bowley, A.L. 1926. *Elements of statistics, Part II: Applications of mathematics to statistics*, 5th ed., London: King.

Box, J.F. 1978. *R.A. Fisher: The life of a scientist*, New York: Wiley.

Brunt, D. 1917. *The combination of observations*, Cambridge: Cambridge University Press.

David, H.A. 1998. 'Statistics in U.S. universities in 1933 and the establishment of the Statistical Laboratory at Iowa State', *Statistical science, 13*, 66–74.

David, H.A. and Edwards, A.W.F. 2001. *Annotated readings in the history of statistics*, New York: Springer.

Dawkins, R. 1986. *The blind watchmaker*, Harlow, UK: Longman.

Edwards, A.W.F. 1972, *Likelihood*, Cambridge University Press.

Edwards, A.W.F. 1990. 'R.A. Fisher: Twice professor of genetics: London and Cambridge *or* "A fairly well-known geneticist" ', *Biometrics, 46*, 897–904.

Fisher, R.A. 1912. 'On an absolute criterion for fitting frequency curves', *Messenger of mathematics, 41*, 155–160.

Fisher, R.A. 1921. 'On the "probable error" of a coefficient of correlation deduced from a small sample', *Metron, 1*, 3–32.

Fisher, R.A. 1922. 'On the mathematical foundations of theoretical statistics', *Philosophical transactions of the Royal Society (A) 222*, 309–368.
Fisher, R.A. 1925. 'Theory of statistical estimation', *Proceedings of the Cambridge Philosophical Society, 22*, 700–725.
Fisher, R.A. 1930a. *The genetical theory of natural selection*, Oxford: Clarendon Press.
Fisher, R.A. 1930b. 'Inverse probability', *Proceedings of the Cambridge Philosophical Society, 26*, 528–535.
Fisher, R.A. 1934. 'Randomisation, and an old enigma of card play', *Mathematical gazette, 18*, 294–297.
Fisher, R.A. 1935. *The design of experiments*, Edinburgh: Oliver and Boyd.
Fisher, R.A. 1936. 'Uncertain inference', *Proceedings of the American Academy of Arts and Sciences, 71*, 245–258.
Fisher, R.A. 1949. *The theory of inbreeding*, Edinburgh: Oliver and Boyd.
Fisher, R.A. 1956. *Statistical methods and scientific inference*, Edinburgh: Oliver and Boyd.
Fisher, R.A. 1971–1974. *Collected papers* (ed. J.H. Bennett), 5 vols., Adelaide: University of Adelaide Press.
Fisher, R.A. 1990. *Statistical methods, Experimental design and scientific inference* (ed. J.H. Bennett), Oxford: Oxford University Press.
Fisher, R.A. and Balmukand, B. 1928. 'The estimation of linkage from the offspring of selfed heterozygotes', *Journal of genetics, 20*, 79–92.
Fisher, R.A. and Yates, F. 1938, *Statistical tables for biological, agricultural and medical research*, Edinburgh: Oliver and Boyd.
Gosset, W.S. 1970. *Letters from W.S. Gosset to R.A. Fisher 1915–1936*, privately printed and circulated. [Foreword by L. McMullen.]
Hald, A. 1998. *A history of mathematical statistics from 1750 to 1930*, New York: Wiley.
Hogben, L.T. 1924. *The pigmentary effector system*, Edinburgh: Oliver and Boyd.
Hogben, L. 1957. *Statistical theory*, London: Allen and Unwin.
Hotelling, H. 1951. 'The impact of R.A. Fisher on statistics', *Journal of the American Statistical Association, 46*, 35–46.
Hoyle, F. 1999. *Mathematics of evolution*, Memphis, Tennessee: Acorn Enterprises LLC.
Kendall, D.G. 1990. 'Obituary: Andrei Nikolaevich Kolmogorov (1903–1987)', *Bulletin of the London Mathematical Society, 22*, 31–100.
Mahalanobis, P.C. 1938. 'Professor Ronald Aylmer Fisher', *Sankhya, 4*, 265–272.
Mather, K. 1951. 'R.A. Fisher's *Statistical methods for research workers*: an appreciation', *Journal of the American Statistical Association, 46*, 51–54.
Pearce, S.C. 1993. 'Introduction to Fisher (1925) Statistical methods for research workers', in S. Kotz and N.L. Johnson (eds.), *Breakthroughs in statistics*, vol. 2, New York: Springer, 59–65.
Pearson, E.S. 1990. *'Student': A statistical biography of William Sealy Gosset*, Oxford: Clarendon Press. [Based on writings by E.S. Pearson, ed. and augmented by R.L. Plackett with G.A. Barnard.]
Snedecor, G.W. 1937. *Statistical methods applied to experiments in agriculture and biology*, Ames, Iowa: Iowa State University Press.
"Student" 1908. 'The probable error of a mean', *Biometrika, 6*, 1–25.
Yates, F. 1951, 'The influence of *Statistical methods for research workers* on the development of the science of statistics', *Journal of the American Statistical Association, 46*, 19–34.
Youden, W.J. 1951. 'The Fisherian revolution in methods of experimentation', *Journal of the American Statistical Association, 46*, 47–50.
Yule, G.U. 1912. *An introduction to the theory of statistics*, 2nd ed., London: Griffin.

CHAPTER 68

GEORGE DAVID BIRKHOFF, *DYNAMICAL SYSTEMS* (1927)

David Aubin

The first book to expound the qualitative theory of systems defined by differential equations, Birkhoff's *Dynamical systems* created a new branch of mathematics separate from its roots in celestial mechanics and making broad use of topology. Important for several fields of mathematics, its impact became massive recently with the spread of 'chaos theory'.

First publication. Providence, Rhode Island: American Mathematical Society (Colloquium Publications Series, no. 9), 1927. viii + 295 pages.

Revised edition. Introduction and addendum by Jürgen Moser, preface by Marston Morse. 1966. xii + 305 pages. Tenth printing 1999.

Russian translation. Dinamicheskie sistemy (trans. E.M. Livenson, ed. A.A. Markov, V.V. Nemytskij and V.V. Stepanov), Moscow and Leningrad: 'Gostekhizdat', 1941. [Repr. Izhevsk: Izd. dom 'Udmurtskij Univ.', Nauchno-Izdatel'skij Tsentr 'Regulyarnaya i Khaoticheskaya Dinamika', 1999 (Seriya Regulyarnaya i Khaoticheskaya Dinamika, no. 8).]

Related articles: Poincaré (§48), Lyapunov (§51), Einstein (§63).

1 INTRODUCTION

'History has responded to these pages on Dynamical Systems in an unmistakable way'. When this book by George David Birkhoff (1884–1944) was reissued in 1966, nearly 40 years after its first publication and more than 20 years after its author's death, Marston Morse stressed its historical legacy in his new preface (p. v). A decade later, such a remark would have seemed superfluous. The craze for 'deterministic chaos' was in full swing and scores of scientists were striving to master dynamical systems theory. Undoubtedly rooted in multifaceted work of Henri Poincaré (1854–1912) at the turn of the century, this theory as Birkhoff defined it was a branch of mathematics that dealt with the global qualitative

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

behavior of systems governed by deterministic laws (that is, where randomness played no part). In retrospect, *Dynamical systems* (hereafter, '*DS*') stands strangely isolated among the mathematical literature of its time as a fundamental intermediary between Poincaré's perceptive work and the modern theory.

Deterministic chaos and dynamical systems theory have had a perplexing history [Aubin and Dahan-Dalmedico, 2002]. That older results that could have been 'forgotten' for several decades gave rise to widespread puzzlement. Albeit well received by the mathematical press when it was first published in 1927, *DS* was a textbook for a field of mathematics that barely existed for some decades to come. Its main domain of application—celestial mechanics—seemed to have lost some of its urgency now that relativity theory and quantum mechanics were revolutionizing physics. By insisting on considering general problems of dynamics as opposed to particular ones and by looking globally at sets of motions rather than particular orbits, Birkhoff's way of approaching the topic was highly original. Not only was he creating an up-to-date topological apparatus for the task at hand, he also confronted head-on the problem of finding a role for dynamical theory when the fundamental equations of physics were being recast. The striking contrast between conformist subject-matter and innovative mathematical and epistemological frameworks can account for the unusual career of *DS*, both the relative oblivion into which it fell and its later success. More than the results presented in the book, the main reason for its posthumous fame is surely its *style*, which largely derives from the intellectual context in which it was produced, that of American mathematics in the decade following the Great War.

2 CELESTIAL MECHANICS: THE HISTORICAL BACKGROUND

In the 19th century, the understanding of the analytic structure of the equations of motion derived from Newtonian mechanics was greatly advanced with the work of Joseph Louis Lagrange (§16), W.R. Hamilton and Carl Jacobi. In the paradigmatic field of celestial mechanics, Pierre-Simon Laplace had perfected Leonhard Euler's perturbation method which allowed him and his successors to compute planetary orbits very accurately in terms of power series (§18). The discovery in 1846 of Neptune on the basis of computations made by Urbain Leverrier and John C. Adams showed that results could be astonishing. But the so-called three-body problem remained as frustrating as ever. The law of gravitation acting upon three masses—especially the Sun, the Earth, and the Moon—gave rise to a system of differential equations for which no explicit expression of the solution valid for all time could be found.

Up to that point, in rational mechanics, one mostly tried to find a local trajectory, that is, solve a system of differential equations with given initial conditions without paying much attention to global behaviors. For complicated problems, the influence of each major planet was treated as a perturbation and solutions expressed in forms of power series. In his famous entry to King Oscar of Sweden's prize of 1889, Poincaré showed that such power series were in general divergent (§48.5). Renouncing the idea of obtaining convergent series, several methods—analytic, extremal or topological—were laid out by Poincaré and explored by a few of his followers. His work on curves defined by differential equations, on celestial mechanics whether concerned with astral orbits or shapes of rotating fluid masses,

even his path-breaking development of topology, always emphasized the global behavior of solutions to a system.

Among the few astronomers and mathematicians who were inspired by Poincaré's ideas, Birkhoff went the furthest in developing a full-fledged theory of dynamical systems detached from its roots in celestial mechanics, and making a systematic use of topology. Having been exposed to the three most prominent currents of mathematical thought in the United States, Birkhoff was well prepared to tread this topological road. His constant insistence on marrying analysis with topology at the highest level of abstraction possible had no other root.

Born in 1884 in Michigan, Birkhoff was awarded a Ph.D. by the University of Chicago in 1907. He was one of the first leading American mathematicians to be fully trained in the Unites States without having made the trip to Europe. But he had divided his time between the leading centers of American mathematics. At Harvard, William Osgood and Maxime Bôcher introduced him to classical analysis, while at Chicago he learnt the abstract modern ideas of Eliakim Hastings Moore's 'general analysis' [Siegmund-Schültze, 1998]. Through his interactions with Oswald Veblen at Princeton University, where he taught from 1909 to 1912, Birkhoff encountered a third significant current of mathematical thought: *analysis situs*, as the nascent field of topology was then called (compare §76.1).

Shortly after Poincaré's untimely death in 1912, Birkhoff suddenly established his international mathematical stature with a *coup d'éclat* when he published the proof of a conjecture known as 'Poincaré's last geometric theorem'. The theorem stated that continuous, one-to-one, area-preserving maps from the annulus to itself rotating points on the boundaries in opposite directions had at least two fixed points [Poincaré, 1912; Birkhoff, *Papers*, vol. 1,673–681]. As Poincaré had already seen, this theorem has important consequences for dynamics.

3 THE CONTENTS OF BIRKHOFF'S BOOK

DS was published in 1927, when Birkhoff was 43 years old; it is summarised in Table 1. A professor of mathematics at Harvard University since 1912, he was by then a well-respected statesman of the American mathematical community, active in the American Academy of Sciences and the National Research Council, as well as having served as the president of the American Mathematical Society.

Before 1927, the only source on general dynamics had been the three volumes of Poincaré's *Les méthodes nouvelles de la mécanique céleste* (1892–1899), characterized by George Darwin as 'for half a century to come [. . .] the mine from which humbler investigators will excavate their materials' [Barrow-Green, 1997, 152]. But Poincaré's magisterial treatise contained much that was cumbersome to use, at times obscure, and at times—for those interested in general dynamics—unduly concerned with details of celestial mechanics. For Birkhoff, on the other hand, dynamics ought not to address a single problem, but rather directly tackle the most general class of dynamical systems defined by the differential equations

$$dx_i/dt = X_i(x_1, \ldots, x_n), \quad i = 1, \ldots, n. \tag{1}$$

Table 1. Contents by chapters of Birkhoff's book.

Ch.	Page	'Title': other included topics
I	1	'Physical aspects of dynamical systems': general analytic discussion, conservation of energy, Lagrangian equations.
II	33	'Variational principles and applications': Hamiltonian dynamics.
III	59	'Formal aspects of dynamics': formal series.
IV	97	'Stability of periodic motions'.
V	123	'Existence of periodic motions': variational principles (geodesics), analytic continuation.
VI	150	'Application of Poincaré's geometric theorem': Poincaré map.
VII	189	'General theory of dynamical systems': topological definitions.
VIII	209	'The case of two degrees of freedom'.
IX	260	'The problem of three bodies'. [End 295.]

In *DS*, Birkhoff summarized more than 15 years of his own research along three main axes: the general theory of dynamical systems; the special case with two degrees of freedom; and the three-body problem in celestial mechanics. These topics form the subject of the last three chapters (VII, VIII, and IX), which have been the most widely admired and studied. In the first two chapters, Birkhoff's treatment was traditional: he gave proofs for existence, uniqueness and continuity theorems, and then discussed Lagrange's equations, Hamiltonian mechanics, and changes of variables. In chapter III, solutions were studied in their formal aspects, that is as power series about which questions of convergence were systematically laid aside as irrelevant to the matter at hand. The next chapter followed Poincaré's idea of investigating the stability of formal solutions near equilibrium or periodic motion. But Birkhoff again went further in considering a vast array of definitions for stability: complete or trigonometric stability, stability of the first order, permanent stability 'for which small displacements from equilibrium remain small over time' (p. 121), semi-permanent stability, unilateral stability (due to Lyapunov: §51), and stability in the sense of Poisson (due to Poincaré).

Chapter V presented four methods by means of which the existence of periodic motions could be established. The first of these made use of the variational principles of dynamics, for example by considering geodesics on a surface and their deformations (a method developed by Jacques Hadamard). A second variational method was called the 'minimax' method, whereby new geodesics were found by considering the lower limit of the length of geodesics stretched by a rotation of the manifold. The minimax method was the original stimulus for Morse theory, which made topological considerations effective for analysis [Morse, 1934, iv]. The third method, due to George W. Hill and Poincaré, looked at the analytic continuation of periodic orbits already known to exist.

The fourth method showing the existence of periodic motion was a generalization of Poincaré's idea of transverse section. Birkhoff's formal theory of Poincaré sections would become an indispensable element of dynamical systems theorists' toolkit (Chapter VI). This gave a dynamical problem a 'striking change of form'. When a continuous dynam-

ical flow was cut transversally by a surface S, each time the continuous dynamical flow was crossing S, the dynamical equation defined a point P_i on the surface. Successive intersections defined a one-to-one analytic transformation T of the section surface S into itself. Poincaré had used the idea to transform the reduced three-body problem into the transformation of the ring into itself. Birkhoff showed that there was a great variety of circumstances where one could use this method. The example of a billiard ball rolling on a flat surface with curved boundaries concretely illustrated the power of the method.

4 BIRKHOFF'S MAIN AMBITION

This was to develop a 'General theory of dynamical systems', which is summarized in Chapter VII of *DS*. 'The final aim of the theory of motion must be directed toward the qualitative determination of all possible types of motions and of the interrelation of these motions' (p. 189). As Koopman [1930] wrote in a review, '[t]he only property made use of in this chapter is the bare fact that the curves are integral curves of analytic differential equations. The treatment has the aspect of a study in point-set theory'. With this work begun in 1912, dynamical systems theory was thoroughly infused with topological ideas.

Birkhoff showed that for arbitrary dynamical systems there always was a closed set of 'central motions' endowed with a certain property of 'recurrence' and towards which all other motions of the system in general tended asymptotically. Considering the equation of motion (1), he looked at states of motion as points in a closed n-dimensional manifold M. To each point P_0 of M (initial conditions), one could associate via (1) a curve of motion lying on M. He then divided the manifold M into two non-intersecting sets: the open set of wandering points, that is, those starting from which the equations of motion would define a trajectory filling open n-dimensional continua in M; and the complementary closed set M_1 of non-wandering points. As time increased or decreased, he showed, every wandering point approached the set M_1 of non-wandering points.

Further, Birkhoff constructed a sequence of sets $M_1, M_2, \ldots$, where M_2 was the set of non-wandering points with respect to M_1, etc. This process had to end at some point with a set C of *central motions.* This was a generalization of periodic motions to which Poincaré had drawn attention. For Birkhoff, the first problem concerning the properties of dynamical systems was the determination of central motions. The fact that for classical dynamics, central motions were the totality of all motion made amply clear that he was constructing a 'general theory' with a wider range of applicability.

In 1912, Birkhoff also introduced the notions of 'minimal' or 'recurrent' sets of motions. Let α- and *ω-limit points* be points towards which other motions tended as time t approached $-\infty$ or $+\infty$. If Σ was a closed, connected set of limit motions (i.e. trajectories composed of limit points of a motion) and Σ had no proper subset, then Birkhoff defined the members of Σ as recurrent motions and the set itself as minimal. He showed that a motion was recurrent if and only if for any $\varepsilon > 0$, curves of motion would remain for a certain interval of time T within a distance ε of every point of the trajectory. In other word, such motions came back arbitrary close to every point of the curve of motion. They were in the set of central motions but the reverse was not necessarily true.

Based on an astute use of topology, these definitions greatly extended the possibility of classifying the motions generated by dynamical systems. Solomon Lefschetz asked in

his review [1929] how classical dynamics was ever able to do without them. Birkhoff's notion of non-wandering points was picked up by the Russian mathematician Aleksandr Andronov, whose roughly contemporary work in this domain ranked equal in importance with Birkhoff's. And it prefigured the concept of 'attractor' that was fundamental for the reconfiguration of dynamical systems theory from the mid-1960s onward. Made famous by René Thom and Steve Smale, an attractor has been succinctly defined as 'an *indecomposable, closed, invariant* set [...] which attracts all orbits starting at points in some neighborhood' [Holmes, 1990].

5 BIRKHOFF'S LECTURE COURSE AND ITS CONTEXT

Though widely admired, the offshoot of the theory was somewhat disappointing. As Birkhoff acknowledged, the remarkable diversity and complexity of behaviors meant that 'rigorously proven qualitative results are rare' [Birkhoff, *Papers*, vol. 2, 246]. As was pointed out by Koopman [1930], what was at stake was the value of a mathematical theory. To reach a better grasp of both Birkhoff's approach to dynamics and the long-term reception of *DS*, one must look at the American postwar context in which it was produced.

Prior to becoming the cornerstone of a branch of mathematics, the American Mathematical Society summer colloquium lectures upon which *DS* was based was the best attended series so far. In September 1920, over 90 mathematicians gathered at the University of Chicago to hear Birkhoff, at the first of these events to take place after the end of the Great War. For many, this was an inspiring return to normalcy. He treated his audience with a review of recent developments in an honored field of mathematical physics crowned with far-ranging philosophical speculations. As was the tradition, the lecturer emphasized his own contributions. In his five lectures, he reviewed traditional approaches to dynamical problems, and summarised his own work on topological tools to describe the various types of motion that could occur in a dynamical system making the crucial distinction between hyperbolic and elliptic motions. His fourth lecture was an application of this 'General analysis' to the three-body problem. Concerned with the 'significance of dynamical systems for general scientific thought', Birkhoff's fifth lecture was not published in *DS*. From their titles, themes broached—'The dynamical model in physics', 'Modern cosmogony and dynamics', 'Dynamics and biological thought', 'Dynamics and philosophical speculation'—are tantalizing in their ambitions [Hurwitz, 1920].

Like other American scientists, mathematicians in the early 20th century were preoccupied by issues of purity [Parshall and Rowe, 1994]: 'Gross utilitarianism is the obvious danger' [Carmichael, 1919, 163]. While astronomers importantly shaped the emerging mathematical community in the United States, a younger generation centered around Chicago 'played a leadership role in defining the mathematical profession on American shores in terms of pure, abstract, rigorous mathematics' [Parshall, 2000, 8]. Their ethos was 'a privileging of pure over applied mathematics, of research over teaching, and of educating future mathematicians over training others who needed advanced mathematical skills' [Butler Feffer, 1997, 66–67]. Birkhoff certainly agreed with a mild version of this credo. Although he always emphasised applications in celestial mechanics, he never computed an orbit.

Since Poincaré had written his *Méthodes nouvelles,* two things had happened that transformed the way dynamics was to be understood: the Great War and relativity theory (§63). If its impact was no way overwhelming, the war effort in 1917–1918 introduced a crucial inflexion in scientists' self-perception [Aubin and Bret, 2003]. A handful of mathematicians, Birkhoff among them, worked on war-related topics (ballistics, sound-ranging, and submarine detection). But as opposed to physicists and chemists, they felt they had a harder time convincing the country that their skills were required for warfare. The war led to a reevaluation of the role played by formal mathematics in the physical sciences and some soul-searching on the mathematicians' part [Servos, 1986]. Some felt that overemphasis on purity had led to a detrimental neglect of applied mathematics. To those concerned with the role of mathematics in science and other human affairs, mathematicians often replied that their inquiries were essential in understanding the deep structures of scientific thought. 'Transcending the flux of the sensuous universe, there exists a stable world of pure thought, a divinely ordered world of ideas, accessible to man, free from the mad dance of time, infinite and eternal' [Keyser, 1915, 679]. Looking for stability in complex flows, dynamical systems theory was Birkhoff's attempt at accommodating two strong, yet antagonistic tendencies of postwar American mathematics: the strive towards purity, if not purism; and the acknowledgment, reinforced by the war, that mathematicians ought to be concerned with applications.

The postwar situation was further complicated by revolutions in physics. Poincaré, it was claimed, 'was depressed when certain recent physical theories seemed to imply that differential equations are not so fundamental to the understanding of phenomena as he had supposed' [Carmichael, 1917, 168]. More than physicists and astronomers, American mathematicians often readily welcomed relativity theory [Goldberg, 1987]. Birkhoff published two books on Einstein's theory (1923, 1925); the first was, with Stanley Eddington's *Mathematical theory of relativity* of 1923, among the earliest books in English explaining relativity with sufficient mathematical sophistication.

Like Veblen, Birkhoff argued that crises in fundamental physics increased the importance of the mathematician who provided a 'rigorous and qualitative background' to the 'more physical, formal, and computational aspects of the sciences' [Birkhoff, *Papers*, vol. 2, 110]. Through the years, his position evolved and it later seemed that dynamical systems theory was for skeptics. 'At a time when no physical theory can properly be termed fundamental—the known theories appear to be merely more or less fundamental in certain directions—it may be asserted with confidence that ordinary differential equations in the real domain, and particularly equations of dynamical origin, will continue to hold a position of the highest importance' (*DS*, iii). 'In view of the many indignities which mechanics has suffered in recent years', a reviewer wrote with a sigh of relief, 'this volume merely illustrates that additional hypotheses are not as yet needed if one wishes to make new discoveries in dynamics' [Bartky, 1928].

6 ON THE IMPACT AND RENAISSANCE OF THE BOOK

DS was not Birkhoff's last word on the topic. In particular, his proof in [Birkhoff, 1931] of the ergodic theorem was deemed as important as his proof of Poincaré's geometric theorem. Introduced by Ludwig Boltzmann, ergodicity has been a cornerstone of statistical

mechanics. It described systems such that each particular motion when continued indefinitely passed through every configuration compatible with energy conservation. Allying topological consideration with Henri Lebesgue's theory of integration (§59.3), Birkhoff developed the notion of transitivity introduced in *DS* (that is, the property of a dynamical system whereby small neighborhoods of curves of motion filled the whole manifold up to a set of measure zero) and showed that it was a widespread property for Hamiltonian systems. Birkhoff knew that this property was not generic, but his results prompted further developments in ergodic theory [Dahan-Dalmedico, 1995].

DS also shaped much of the work done by Aleksandr Kolmogorov, Vladimir I. Arnol'd, and Jürgen Moser in the 1950s and 1960s on the celebrated KAM theorem that invalidated Birkhoff's ergodic conjecture [Diacu and Holmes, 1996]. Several other concepts introduced by Birkhoff were later picked up by others. On the notion of recurrent motion, Morse, Walter H. Gottschalk, and Gustav A. Hedlund built an abstract theory of symbolic dynamics in 1955 that is used today in theoretical computer science. Another example is the 'bad' curve studied by Birkhoff in 1932, a complicated state of motion that ultimately formed the basis for Smale's 'horseshoe', a stable, yet chaotic motion [Abraham, 1985].

Very technical, those developments kept the memory of Birkhoff's *DS* alive, but restricted to specialized fields of inquiry until dynamical systems theory was spectacularly revived after the Second World War by Lefschetz [Dahan-Dalmedico, 1994]. But Lefschetz and his collaborators rediscovered the work of Poincaré through their close study of Russian sources rather than in Birkhoff's work. One reason for this was the insistence put on dissipative systems where energy is not conserved, as opposed to conservative ones emphasized by Birkhoff. In *DS*, the section on dissipative systems occupied less than two pages. He acknowledged that '[c]onservative systems are often limiting cases of what is found in nature', but dissipative systems generally tended toward a the motion of a conservative system with fewer degrees of freedom.

The mathematicians' more active participation to the Second World War and the Cold War, as well as concerns with nonlinear oscillations arising from radio-engineering (B. Van der Pol), led to an understanding of dynamical systems different than that stemming from Birkhoff's nearly exclusive concern with celestial mechanics. This crucial difference in emphasis is brought to light by comparing Birkhoff's attitude concerning stability with Andronov's [Dahan-Dalmedico, 2004]. Both dealt with general systems of differential equations using many of the same sources (Poincaré, Lyapunov). They nonetheless ended up with almost opposite views on stability. Inspired by the famous 'problem of stability' of the three-body problem, Birkhoff restricted the study of stability to that of orbits lying near a periodic (or central) motion. He thought one had to dictate, by convention or by a judicious choice of problems to be answered, the kind of stability that one wanted to look at. 'All that stability can mean is that, for the system under consideration, those motions whose curves lie in a certain selected part of phase space from and after a certain instant are *by definition* called stable, and other motions unstable' [Birkhoff, *Papers*, vol. 3, 602]. Concerned with radio systems, Andronov imagined a more general type of stability that applied not only to solutions of a system of differential equations, but to the system itself. The only interesting systems for modeling, he thought, were structurally stable, that is, keeping the same qualitative behavior under small deformations.

In the 1960s, a generation too young to have had to participate to the war effort revived widespread interest in *DS*. Less interested in control than their elders, Mauricio Peixoto and Smale launched a general program of classification of dynamical systems that thrived on the topological approach that was characteristic of Birkhoff's work. Following the publication of Smale's *Differentiable dynamical systems* in 1967, a blooming field was established that had a profound impact on the way that the mathematical modeling of natural phenomena was to be understood. Edward N. Lorenz in 1963 and David Ruelle in 1971 independently exhibited systems governed by simple deterministic laws that nonetheless exhibited complex, apparently erratic behaviors [Aubin, 2001].

All of a sudden, *Dynamical systems* enjoyed a second life. In this book, people interested in chaos found a straightforward style that corresponded to their expectations. Physicists liked to see equations of motion written in a form they recognized. They were comfortable with discussions of Lagrangian and Hamiltonian functions. No fancy Bourbakist abstraction here defaced them [Aubin, 1997]. No more than an elementary topological knowledge was required to grasp the most innovative ideas introduced in the book. Readers also appreciated the self-contained character of the book and the tools presented in all generality, in less than 300 pages of clear English prose. A generation mobilized against the Vietnam war and intend to 'explicitly direct [its] work toward socially-positive goals' [Smale, 1972, 3] found in American struggles with issues of purity after the Great War in the face of new wars and upheavals in physics an epistemological and moral framework with which they felt comfortable.

BIBLIOGRAPHY

Reviews of the book are gathered at the end.

Abraham, R.H. 1985. 'In pursuit of Birkhoff's chaotic attractor', in S.N. Pnevmatikos (ed.), *Singularities and dynamical systems*, Amsterdam: North-Holland, 303–312.

Archibald, R.C. 1938. *A semicentennial history of the American Mathematical Society, 1888–1938*, New York: American Mathematical Society. [Repr. 1988.]

Aubin, D. 1997. 'The withering immortality of Nicolas Bourbaki: a cultural connector at the confluence of mathematics, structuralism, and the Oulipo in France', *Science in context, 10*, 297–342.

Aubin, D. 2001. 'From catastrophe to chaos: the modeling practices of applied topologists', in A. Dahan-Dalmedico and U. Bottazzini (eds.), *Changing Images in mathematics: From the French Revolution to the new millennium*, London: Routledge, 255–279.

Aubin, D. and Bret, P. (eds.) 2003. *Le sabre et l'éprouvette: l'invention d'une science de guerre 1914–1939*, Paris: Éditions Noesis/Agnès Viénot, «14–18» n° 6.

Aubin, D. and Dahan-Dalmedico, A. 2002. 'Writing the history of dynamical systems and chaos: *longue durée* and revolution, disciplines and cultures', *Historia mathematica, 29*, 273–339.

Barrow-Green, J. 1997. *Poincaré and the three body problem*, Providence: American Mathematical Society; London: London Mathematical Society.

Birkhoff, G. 1989. 'Mathematics at Harvard, 1836–1944', in [Duren, 1989], 3–58.

Birkhoff, G.D. *Papers*. *Collected mathematical papers*, 3 vols., Providence: American Mathematical Society, 1950. [Repr. New York: Dover, 1968.]

Birkhoff, G.D. 1925. *The origin, nature, and influence of relativity*, New York: Macmillan.

Birkhoff, G.D. 1931. 'Proof of the ergodic theorem', *Proceedings of the National Academy of Science, 17*, 656–660. [Repr. in *Papers*, vol. 2, 404–408.]

Birkhoff, G.D. with Langer, R.E. 1923. *Relativity and modern physics*, Cambridge, MA: Harvard University Press.

Butler Feffer, L. 1997. 'Mathematical physics and the planning of American mathematics: ideology and institutions', *Historia mathematica, 24*, 66–85.

Carmichael, R.D. 1917. 'The provision made by mathematics for the needs of science', *Science, 45*, 465–474.

Carmichael, R.D. 1919. 'Motives for the cultivation of mathematics', *Scientific monthly, 8*, 160–178.

Dahan-Dalmedico, A. 1994. 'La renaissance des systèmes dynamiques aux États-Unis après la deuxième guerre mondiale: l'action de Solomon Lefschetz', *Supplemento ai Rendiconti del Circolo Matematico di Palermo*, ser. 2, no. 34, 133–166.

Dahan-Dalmedico, A. 1995. 'Le difficile héritage de Henri Poincaré en systèmes dynamiques', in *Sonderdruck aus Henri Poincaré: Science et philosophie, Congrès international de Nice, 1994*, Berlin: Akademie Verlag; Paris: Albert Blanchard, 13–33.

Dahan-Dalmedico, A. with Gouzevitch, I. 2004. 'Early developments of nonlinear science in Soviet Russia: The Andronov School at Gor'kiy', *Science in context, 17*, 235–265.

Diacu, F., and Holmes, P.J. 1996. *Celestial encounters: the origins of chaos and stability*, Princeton: Princeton University Press.

Duren, P. and others (eds.) 1989. *A century of mathematics in America*, pt. 2, Providence: American Mathematical Society.

Goldberg, S. 1987. 'Putting new wine in old bottles: the assimilation of relativity in America', in T.F. Glick (ed.), *The comparative reception of relativity*, Dordrecht and Boston: Reidel, 1–26.

Hedrick, E.R. 1917. 'The significance of mathematics', *Science, 46*, 395–399.

Holmes, P.J. 1990. 'Poincaré, celestial mechanics, dynamical-systems theory, and "chaos" ', *Physics reports, 193*, 137–163.

Hurwitz, W.A. 1920. 'The Chicago Colloquium', *Bulletin of the American Mathematical Society, 27*, 65–71.

Kellogg, O.D. 1921. 'A decade of American mathematics', *Science, 53*, 541–548.

Keyser, C.J. 1915. 'The human significance of mathematics', *Science, 42*, 663–680.

Miller, G.A. 1917. 'The function of mathematics in scientific research', *Science, 45*, 549–558.

Morse, M. 1934. *The calculus of variations in the large*, Providence: American Mathematical Society.

Morse, M. 1946. 'George David Birkhoff and his mathematical work', *Bulletin of the American Mathematical Society, 52*, 357–391. [Repr. in [Birkhoff, *Papers*], vol. 1, xxiii–lvii (cited here).]

Parshall, K.H. 2000. 'Perspectives on American mathematics', *Bulletin of the American Mathematical Society, n.s. 37*, 381–405.

Parshall, K.H. and Rowe, D.E. 1994. *The emergence of the American Mathematical research community, 1876–1900: J.J. Sylvester, Felix Klein, and E.H. Moore*, Providence: American Mathematical Society; London: London Mathematical Society.

Poincaré, H. 1912. 'Sur un théorème de géométrie', *Rendiconti del Circolo Matematico di Palermo, 33*, 375–407. [Repr. in *Oeuvres*, vol. 6, 499–538.]

Rothrock, D.A. 1919. 'Mathematicians in War service', *American mathematical monthly, 26*, 40–44. [Repr. in [Duren, 1989], 269–273.]

Servos, J.W. 1986. 'Mathematics and the physical sciences in America, 1880–1930', *Isis, 77*, 611–629.

Siegmund-Schültze, R. 1998. 'Eliakim Hastings Moore's general analysis', *Archive for history of exact sciences, 52*, 51–89.

Smale, S. 1972. 'Personal perspectives on mathematics and mechanics', in S.A. Rice, K.T. Freed and J.C. Light (eds.), *Statistical mechanics: new concepts, new problems, new applications*, Chicago: University of Chicago Press, 3–12.

Veblen, O. 1946. 'George David Birkhoff (1884–1944)', *Yearbook of the American Philosophical Society*, 279–285. [Repr. in [Birkhoff, *Papers*], vol. 1, xv–xxiii; also in *Biographical memoirs of the National Academy of Sciences*, *80* (2000), 44–57.]
Whittaker, E.T. 1945. 'George David Birkhoff', *Journal of the London Mathematical Society*, *20*, 121–128.

Reviews of Dynamical systems consulted:

Bartky, W. 1928. *American mathematical monthly, 35*, 561–563.
Buhl, A. 1928. *L'Enseignement mathématique, 27*, 170–171.
Cherry, T.M. 1928–1929. *Mathematical gazette, 14*, 198–199.
Koopman, B.O. 1930. *Bulletin of the American Mathematical Society, 36*, 162–166.
Lefschetz, S. 1929. *Bulletin des sciences mathématiques, 53*, 193–195.

CHAPTER 69

P.A.M. DIRAC (1930) AND J. VON NEUMANN (1932), BOOKS ON QUANTUM MECHANICS

Laurie M. Brown and Helmut Rechenberg

Dirac's book is the classic physicist's treatise on the quantum mechanical transformation theory, which is a generalization of the matrix-mechanical and wave-mechanical quantum theories of Werner Heisenberg and Erwin Schrödinger, respectively. The book of von Neumann covers the foundational principles of quantum mechanics in a careful mathematical way.

Dirac, The principles of quantum mechanics, first edition

First publication. Oxford: Clarendon Press, 1930. vii + 257 pages.

Later editions. 2nd 1935, 3rd 1947, 4th 1958, and '4th revised' (1967; repr. 1971, 1974, 1981 and later): all Clarendon Press.

Translations. In many languages. Due to Dirac's relationship with Peter Kapitza, noteworthy are the three Russian editions, with publishers' prefaces having ideological content [Dirac, 1995, 472–478]; *Osnovi kvantoboy mekhaniki* (trans. M.P. Bronshtein), 1st ed., Moscow and Leningrad: State Technico-theoretical Publishing, 1932. Also the Japanese transl. of the 2nd and subsequent editions, by Y. Nishina, S. Tomonaga, M. Kobayashi, and H. Tamaki.

von Neumann, Mathematische Grundlagen der Quantenmechanik

First publication. Berlin: Julius Springer, 1932. ii + 262 pages.

English translation. Mathematical foundations of quantum mechanics (trans. R.T. Beyer), Princeton: Princeton University Press, 1955. xii + 455 pages.

Related articles: Kelvin (§58), Lorentz (§60).

1 THE DISCOVERY OF QUANTUM MECHANICS

Beginning in 1913, Niels Bohr (1885–1962) and Arnold Sommerfeld (1868–1951) had applied the quantum-theoretical ideas of Max Planck and Albert Einstein to the nuclear

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

atomic model of Ernest Rutherford and created what became the 'old quantum theory'. This theory, which was based closely upon classical mechanics with certain restrictions, had achieved some successes [Sommerfeld, 1919], but since 1923 its complete failure to deal with the spectra of many-electron atoms led to efforts to replace it. In July 1925 Werner Heisenberg (1901–1976) in Göttingen, a former student of Sommerfeld and assistant to Max Born (1882–1970), completed an article in which he proposed a radically new description of periodic atomic systems, using what he called 'quantum theoretically reformulated Fourier series' involving only observable quantities, namely, transition amplitudes and frequencies. The dynamical variables defined in this scheme did not obey the commutative law of multiplication [Heisenberg, 1925]. He showed the article to Born, who realized that Heisenberg's new variables were matrices. Together with Heisenberg and Pascual Jordan (1902–1980), the three developed the first mathematical theory of quantum mechanics, known as *matrix mechanics* [Born and Jordan, 1930]. In August 1925 Heisenberg sent a copy of the proof sheets of his pioneering paper to England, where Paul Dirac (1902–1984) studied it and formulated an alternative mathematical theory of quantum mechanics, known as *quantum algebra*.

At the request of David Hilbert (1862–1943), in fall 1925 Heisenberg presented a survey of the new quantum mechanics in the Göttingen mathematical seminar. Starting from empirical quantum phenomena, he showed that the dynamical variables could be represented by infinite Hermitean matrices and demonstrated how their eigenvalues (denoting the empirical data) were obtained with the help of the well-known method of principal-axis transformation. Hilbert had introduced this method earlier (1904–1910) in connection with his theory of linear integral equations, where he had also shown the equivalence of discrete matrix and continuous integral-equation representations of the eigenvalue problem [Hilbert, 1912]. When Erwin Schrödinger (1887–1961) presented his wave mechanics in January 1926 it appeared, therefore, obvious that his treatment of atomic problems with differential equations, the matrix method of Born and others, and the algebraic method of Dirac were equivalent.

Although then of feeble health, Hilbert was very interested in a strict mathematical formulation of quantum mechanics. In the winter semester 1926–1927 he delivered a lecture course on 'Mathematical methods of quantum theory', in which he stressed the necessity to develop an axiomatic basis for the *statistical transformation theory*, which had been proposed in December 1926 by Dirac and Jordan. Hilbert's guest John von Neumann (1903–1957) and his assistant Lothar Nordheim prepared a publication of excerpts from his lectures [Hilbert et alii, 1928]. Von Neumann would then carry out Hilbert's program in a set of papers on the mathematical foundation of quantum mechanics (1927–1930), which were extended and summarized in his book of 1932.

2 BACKGROUND OF DIRAC'S *PRINCIPLES*

Heisenberg's new theory was meant to replace the 'old quantum mechanics' of Bohr and Sommerfeld, which had proven to be unsatisfactory. Heisenberg sent a proof copy of the article (received by the journal on 29 July 1925) to Ralph Fowler at Cambridge University, who passed it on to his research student Dirac with the note: 'What do you think of this? I shall be glad to hear. RHF'.

On 20 July, Born asked his student Jordan to help him to extend Heisenberg's work. During the next few months, Born and Jordan, soon joined by Heisenberg, wrote two papers that practically completed Heisenberg's ideas in the form of *matrix mechanics* [Born and Jordan, 1925; Born, Heisenberg, and Jordan, 1926]. Wolfgang Pauli, who had been following these developments closely, promptly obtained the spectrum of the hydrogen atom, using Heisenberg's 'new quantum mechanics', and pointed out that it avoided difficulties in the old theory that arose when crossed electric and magnetic fields were present.

Meanwhile, independently of the Göttingen group, Dirac at Cambridge obtained essentially the same results as Born and Jordan, as Dirac described in an interview [Van der Waerden, 1967, 41]:

> At first I could not make much of [Heisenberg's paper], but after about two weeks I saw that it provided the key to the problem of quantum mechanics. I proceeded to work it out by myself. I had previously learnt the Transformation Theory of Hamiltonian Mechanics from lectures by R.H. Fowler and from Sommerfeld's book Atombau und Spektrallinien.

One of the main results of Born and Jordan was that if q and p are quantum variables corresponding to a pair of canonically conjugate variables in the classical Hamiltonian theory, then their commutator, defined by $qp - pq$, has the value $i\hbar$, where $\hbar = h/2\pi$ and h is Planck's constant. (For example, if q is the Cartesian coordinate x, then p is the momentum p_x.) On the other hand, quantum coordinates and momenta that are not canonically conjugate do commute.

In his paper Dirac [1925] obtained the same result. However, he also observed that the commutator was analogous to a general expression, called the Poisson bracket, that occurs in classical Hamiltonian mechanics. This is defined as follows: Let q_r and p_r (with $r = 1, 2, 3, \ldots$) be a complete set of canonically conjugate variables, and let x and y be functions of these variables. Then, the Poisson bracket of x and y is

$$[x, y] = \sum_r \left\{ \frac{\partial x}{\partial q_r}\frac{\partial y}{\partial p_r} - \frac{\partial y}{\partial q_r}\frac{\partial x}{\partial p_r} \right\}. \tag{1}$$

Using Bohr's Correspondence Principle, which states that for large quantum numbers classical physics should apply, Dirac conjectured that the Poisson bracket should correspond to the quantum commutator, or more precisely, in the large quantum number limit, that

$$xy - yx = i\hbar[x, y]. \tag{2}$$

Dirac had discovered a way to translate the equations of Hamiltonian dynamics into the language of quantum mechanics. For example, the Hamiltonian equations of motion are ($\dot{q}_r = dq_r/dt$, etc. and $H(q_r, p_r)$ is the Hamiltonian function):

$$\dot{q}_r = [q_r, H], \qquad \dot{p}_r = [p_r, H], \tag{3}$$

and more generally, for any function $u(q_r, p_r)$,

$$du/dt = [u, H] + \partial u/\partial t. \tag{4}$$

These equations are also valid as quantum mechanical equations, provided one replaces the Poisson bracket by the commutator, as in (2).

In a second paper, Dirac [1926a] generalized the matrix formulation of quantum mechanics, using a more abstract algebraic scheme. In this system, he called the non-commuting quantum variables *q-numbers*, distinguishing them from the classical commuting quantities that he called *c-numbers*. He stated that the quantum calculations are to be carried out in terms of the q-numbers. However, the results of the calculations have then to be interpreted as c-numbers, in order to be compared with experiment. He called his new method *transformation theory* [Mehra and Rechenberg, 1982; Kragh, 1990].

In addition to the articles of Pauli and Dirac, the journals received another major work in January 1926. That was the first of a series of papers, in which Schrödinger introduced the quantum wave function and wrote the famous equation for it that bears his name [Schrödinger, 1926a]. In March he showed the equivalence of his new formulation, expressed in terms of eigenfunctions and eigenvalues of linear operators, to that of Heisenberg, Born, and Jordan [Schrödinger, 1926b]. In a subsequent paper, Schrödinger [1926c] became the first person to calculate the intensities of the lines of the atomic hydrogen spectrum. Dirac was very pleased to discover that he could generalize his transformation theory to include both wave mechanics and matrix mechanics, and thus he provided a second proof of their equivalence [Dirac, 1927a]. The Schrödinger and Heisenberg 'pictures' are limiting cases, and other intermediate representations are possible as well.

Born in Bristol, England on 8 August 1902, Paul Dirac studied there at the Merchant Venturers Technical College, where his father was a teacher of French. After three years study of electrical engineering and two years of applied mathematics, in 1923 he became a research student at St. John's College, Cambridge, with Ralph Fowler as his adviser. In 1926, Dirac received his doctorate from Cambridge University with a dissertation entitled 'Quantum mechanics'. From 1926 to 1930, Dirac made other major contributions to quantum theory, including quantum statistics, quantum electrodynamics, and the relativistic quantum theory of the electron. For their pioneering work on quantum mechanics, Heisenberg, Dirac and Schrödinger were all awarded in 1933 the Nobel Prize in Physics. The 1932 prize, not awarded in that year, went to Heisenberg, while Dirac and Schrödinger shared the prize for 1933.

With regard to quantum statistics, Albert Einstein and Satyendra Nath Bose had in 1924 introduced the so-called Einstein–Bose (E–B) statistics, which apply to identical particles of integer spin, such as photons and He^4 atoms. Enrico Fermi had introduced in 1926 the statistics that are responsible for Pauli's exclusion principle. They apply to electrons and other particles of half-integral spin, such as He^3 atoms and are referred to as Fermi–Dirac (F–D) statistics. Dirac's contribution [Dirac, 1926b] was to show that the two kinds of statistics correspond to Schrödinger wave functions that are, respectively, symmetric (E–B case) or antisymmetric (F–D case) under the exchange of identical particles.

In 1927, Dirac published a quantum theory of the electromagnetic field interacting with electrons, the so-called *quantum electrodynamics* or *QED*. Born and Jordan had proposed in 1925 a quantum theory of the free electromagnetic field, which Dirac developed into a practical scheme for calculating rates of emission and absorption of radiation [Dirac, 1927b]. Making a Fourier decomposition of the classical electromagnetic field, he regarded

each frequency component as arising from a harmonic oscillator. Dirac then replaced the classical oscillators by quantum ones and interpreted the nth quantum state of an oscillator of angular frequency ω to represent n light quanta of energy $\hbar\omega$. In this way an assembly of photons replaces a train of waves. This mathematical equivalence was the resolution of the wave-particle paradox that had puzzled physicists for decades.

Perhaps Dirac's greatest achievement was a relativistic quantum theory of the electroni [Dirac, 1928]. Among other advances, this theory soon predicted the existence of an entirely unexpected new form of matter, called antimatter. Although there already existed a relativistic Schrödinger equation, it did not include the effects of electron spin and thus did not provide an accurate spectrum of atomic hydrogen. This equation, also called the Klein–Gordon (KG) equation, was a partial differential equation of second order in the time, analogous to the free particle relation

$$E^2 = c^2\boldsymbol{p}^2 + m^2c^4, \tag{5}$$

where $E =$ energy, $\boldsymbol{p} = (p_x, p_y, p_z) =$ momentum, $m =$ mass and $c =$ speed of light. The KG equation did not conform to Dirac's transformation theory, since Hamiltonian dynamics has equations of motion that are linear in time.

Dirac regarded this as a major shortcoming and tried to deal with it by writing a Hamiltonian function in the linear form

$$H = \boldsymbol{\alpha}\cdot\boldsymbol{p} + \beta m. \tag{6}$$

Since H represents the relativistic energy, squaring both sides of (6) should resemble (5). That is not possible if $\boldsymbol{\alpha} = (\alpha_x, \alpha_y, \alpha_z)$ and β are numbers, but it can work if they are suitable anticommuting 4×4 matrices. This results in the famous Dirac equation, which we can write in a convenient notation as:

$$\Pi_0\psi = (\boldsymbol{\alpha}\cdot\boldsymbol{\Pi} + \alpha_0 m)\psi, \quad \text{with } \alpha_i\alpha_j + \alpha_j\alpha_i = 2\delta_{ij},\ i, j = 0, 1, 2, 3. \tag{7}$$

Here the Dirac wave-function ψ has four components $(\psi_1, \psi_2, \psi_3, \psi_4)$, the α_i are 4×4 matrices, and we have set $\beta = \alpha_0$. We also replaced $\boldsymbol{p} = -i\hbar\nabla$ and $H = p_0 = i\hbar\partial/\partial t$, respectively, by $\boldsymbol{\Pi} = \boldsymbol{p} - (e/c)\boldsymbol{A}$, and $\boldsymbol{\Pi}_0 = p_0 - (e/c)A_0$, where A_μ $(\mu = 0, 1, 2, 3)$ is the electromagnetic four-vector potential.

The Dirac equation (7) gave excellent agreement with the hydrogen spectrum and with other experiments, including the scattering of high-energy radiation on electrons (Compton scattering). It automatically gave the electron spin of $1/2\hbar$ and the magnitude of its magnetic moment $e\hbar/mc$. However, its successes were accompanied by a very disturbing feature: The Dirac equation implied that electrons could have *negative* total energy, which would imply negative mass, from Einstein's principle that $E = mc^2$. As we will relate below, several years were needed to discover the true interpretation of those negative energy solutions.

3 THE *PRINCIPLES OF QUANTUM MECHANICS*

In 1930, Dirac published the landmark treatise called *The principles of quantum mechanics*; its contents are summarised in Table 1. It has been justly compared with the *Principia*,

Table 1. Contents by chapters of Dirac's book

Chapter	Page	Title
I	1	The principle of superposition.
II	18	Symbolic algebra of states and observables.
III	35	Eigenvalues and eigenstates.
IV	55	Representations of states and observables.
V	73	Transformation theory.
VI	92	Equations of motion and quantum conditions.
VII	117	Elementary applications.
VIII	137	Motion in a central field of force.
IX	157	Perturbation theory.
X	176	Collision problems.
XI	198	Systems containing several similar particles.
XII	218	Theory of radiation.
XIII	238	Relativity theory of the electron. [End 257.]

the masterwork of one of Dirac's predecessors in the Lucasian Chair of Mathematics at Cambridge University, namely Isaac Newton. Dirac set out his program of the *Principles* in the Preface to the first edition, which is reprinted almost in its entirety in all subsequent editions. Contrasting the new theory of quantum mechanics with the classical tradition, where 'we could form a mental picture in space and time of the whole scheme', he pointed out that 'nature works on a different plan'. Dirac continued:

> Here fundamental laws do not govern the world as it appears in our mental picture in any very direct way, but instead they control a substratum of which we cannot form a mental picture without introducing irrelevancies. The formulation of these laws requires the use of the mathematics of transformations [. . .]. The growth of the use of transformation theory, as applied first to relativity and later to the quantum theory, is the essence of the new method in theoretical physics.

For this reason, he argued that a book on the new physics must be essentially mathematical. However, he stated:

> All the same the mathematics is only a tool and one should learn to hold the physical ideas in one's mind without reference to the mathematical form. In this book I have tried to keep the physics to the forefront, by beginning with an entirely physical chapter and in the later work examining the physical meaning underlying the formalism wherever possible.

Indeed, in the first edition of *Principles*, the initial chapter has no equations, and in the later editions the first chapter has very few. As for the mathematical style, Dirac pointed out two possibilities, namely: an abstract symbolic method or a method using 'coordinates or

representations'. Except for Hermann Weyl's book *Gruppentheorie und Quantenmechanik* [Weyl, 1928], all other treatments used the latter method, whether it be matrix mechanics or wave mechanics. Dirac adopted the symbolic method, feeling that 'it goes more deeply into the nature of things'.

After remarking that 'it is quite hopeless on the basis of classical ideas to try to account for the remarkable stability of atoms and molecules', Dirac's first chapter considers the *principle of superposition* of states as the main property of the new mechanics. This principle is borrowed from the classical theory of waves, but it is here applied to particles as well. Light, which exhibits such typical wave phenomena as interference and diffraction, also appears as *photons*, energy 'packages' of $E = h\nu$, where ν is the light frequency, as required by Einstein's explanation of the photoelectric effect. Thus light consists of both waves and particles.

Consider a plane-polarized beam of light of very low intensity entering a plane polarizer (such as a crystal), so weak that we can consider one photon at a time. Let the polarizer be set at an angle α to the photon's plane of polarization. As the photon's integrity is always preserved, it will either pass through the polarizer (with probability $\cos^2\alpha$) or be absorbed (with probability $\sin^2\alpha$). We are unable to predict the outcome of the experiment and must be content to know the *probability* of the photon's transmission.

Until the measurement is made, that is, until we observe that the photon has either passed through the polarizer or not, we can regard the photon as being in a mixture of two states, one that is polarized in the plane of the polarizer and one that is polarized in a perpendicular plane. Alternatively, we can use any two perpendicular planes of polarization to form the mixture. This idea can be generalized to the case of an atom that can be in a mixture of atomic states (p. 10):

> When an observation is made on any atomic system that has been prepared in a given way and is thus in a given state, the result will not in general be determinate, *i.e.* if the experiment is repeated several times under identical conditions several different results may be obtained. If the experiment is repeated a large number of times it will be found that each particular result will be obtained a definite fraction of the total number of times, so that one can say there is a definite probability of its being obtained any time that the experiment is performed. This probability the theory enables one to calculate. In special cases this probability may be unity and the result of the experiment is then quite determinate.

This leads to the *principle of superposition*, which is different from the classical notion of superposition of waves (p. 15):

> We may say that a state A may be formed by a superposition of states B and C when, if any observation is made on the system in state A leading to any result, there is a finite probability for the same result being obtained when the same observation is made on the system on one (at least) of the two states B and C. The Principle of Superposition says that any two states B and C may be superimposed in accordance with this definition to form a state A and indeed an infinite number of different states A may be formed by superposing B and C

> in different ways. This principle forms the foundation of quantum mechanics. It is completely opposed to classical ideas, according to which the result of any observation is certain and for any two states there exists an observation that will certainly lead to two different results.

The second chapter of *Principles* introduces the mathematics of quantum mechanics and the remaining first half of the book develops this subject in an abstract way, the applications to physical problems filling out the second half. This scheme is preserved in all the editions. In the first edition Dirac argued (p. vi): 'From the mathematical side the approach to the new theories presents no difficulties, as the mathematics required [...] is not essentially different from what has been current for a considerable time'.

The mathematics may not be 'essentially different' from that in common use by physicists, but Dirac's innovations have provided the stimulus for at least two new branches of mathematics, namely the theory of *distributions* and the theory of the *Dirac operator* (or 'square root of the Laplacian'), used in formulating Dirac's relativistic electron equation.

In the first edition, Dirac constructed an abstract scheme in which 'physical things such as states of a system or dynamical variables' are represented by algebraic symbols that are analyzed in accordance with certain axioms and laws. In the later editions (which physicists considered to be more accessible, or as Heisenberg said '*menschlicher*'), Dirac adopted a geometric picture to represent physical states, employing a space of complex vectors and its dual space; the state vectors have unit norm. Linear operators (q-numbers) act upon the state vectors, transforming the space into itself. When such a space is denumerably infinite, it is called a Hilbert space, introduced originally in David Hilbert's theory of linear integral equations. Quantum mechanics, however, requires a generalization to a continuously infinite-dimensional vector space, and this means new mathematics. Hermann Weyl, a student and collaborator of Hilbert, explicitly discussed this point [Weyl, 1928].

Although the laws of quantum mechanics are expressed in terms of the abstract q-numbers, the application to physical problems demands the use of representations to obtain c-number predictions that can be compared to experiment. To deal with continuous variables like position and momentum, leading to continuously infinite-dimensional vector spaces, Dirac introduced his famous δ-function. The one-dimensional $\delta(x)$ is an improper function that is zero everywhere except at $x = 0$, where it is infinite; its integral over all x is unity. It is clearly a generalization of the Kronecker $\delta_{ij,}$, having integer indices, which is zero for $j \neq i$ and unity for $j = i$. Dirac was aware that the use of $\delta(x)$ was not rigorous mathematics, but he argued that it use 'will not be in itself a source of lack of rigour in the theory, since any equation involving the δ function can be transformed into an equivalent but usually more cumbersome form in which the δ function does not appear'. The mathematician Laurent Schwartz, who invented the rigorous distribution theory in 1945, realized that Dirac had anticipated his work. Later he remarked: 'It was not only the Dirac "function" itself that Dirac put forward, but likewise for all singular functions, he had the idea of distributions as kernels'.

4 LATER EDITIONS OF *PRINCIPLES*

In the preface to the second edition, Dirac says: 'The book has been mostly rewritten'. The theory, developed in a less abstract form, 'should make the work suitable for a wider circle

of readers'. The content was mostly unchanged, but emphasized the geometry of the space of state vectors. In general, reviewers preferred the new pedagogical style.

The chapter headings of the third edition of *Principles* are very similar to those of the first two editions, but in the third and succeeding editions he used a new notation that he developed starting from 1939. According to a review by Herman Feshbach, the effect was 'to render the relation between states and wave functions more transparent, many of the proofs become shorter and clearer'. Also, one could use dyadics in state vector space to represent linear operators.

The new notation can be symbolized as $\langle \text{bra}|\text{c}|\text{ket}\rangle$, where the ket or ket vector $|\text{ket}\rangle$ represents a quantum state. This state can be labeled by one or more parameters, for example, $|a, b, \ldots\rangle$. The bra or bra vector is the *conjugate imaginary* vector; for example, $\langle a, b, \ldots|$ is the conjugate imaginary of $|a, b, \ldots\rangle$ in the sense that the product $\langle a, b, \ldots|a, b, \ldots\rangle$ is a real positive number (note that $|\ |$ has been shortened to $|$). In general $\langle k|l\rangle$ is a number whose complex conjugate is $\langle l|k\rangle$. The c in the expression $\langle \text{bra}|\text{c}|\text{ket}\rangle$ is a linear operator, so that $\text{c}|\text{ket}\rangle$ is also a ket and $\langle \text{bra}|\text{c}$ is also a bra.

The fourth and 'fourth revised' editions are essentially identical to the third edition, except for the final chapters, dealing with the relativistic electron theory and quantum field theory. They appear in the *Principles*, beginning with the first edition and form a very important part of Dirac's work.

Chapter XIII of the first edition deals with Dirac's relativistic electron theory. For simplicity (and because he was not ready to accept a quantum field theory of the electron), he considered a single electron described by a four-component Schrödinger wave function, as in (7) above. He gave the wave function for a free electron and also calculated the spectrum of the hydrogen atom.

He then turned to the troublesome question of the interpretation of the electron states of negative energy. We note that we are speaking here of the states of *total* (not relative) negative energy, whose existence relativistically would imply negative mass. In addition, ordinary electrons could fall into these states (in quantum theory, though not classically, where they would have to traverse a forbidden gap in energy). In order to prevent this catastrophe, Dirac invented his 'hole theory', according to which *almost* all of the negative energy states are occupied and hence, according to the Pauli Exclusion Principle, they are not accessible to positive energy electrons. Dirac assumed that a completely filled set of vacuum states would not appear as electrically charged.

However, a sufficiently energetic photon could lift a negative energy electron to a positive state, leaving behind an observable hole. Dirac argued that this hole in the negative 'sea' would behave as a positive charge, and he identified it with the only known elementary particle, namely, the proton, which is much heavier than the electron. The presentation of all the material on the relativistic electron in the second and all successive editions was almost unchanged, *except* that everywhere the word 'proton' appears it is replaced by *positron*. This change occurred for two reasons. In the first place, several physicists proved that the positively charged 'hole' in Dirac's theory must have exactly the same mass as the electron, the theory being charge-symmetric. In the second place, the positive electron, or positron, was observed in 1932 to be a component of the cosmic rays, and the positron was soon afterwards produced in the laboratory, both as a member of an electron positron pair

(through gamma ray materialization) and as a product of artificially induced nuclear beta decay.

In the first three editions, Dirac presented a version of quantum electrodynamics (hereafter, '*QED*') in which the electromagnetic field is quantized (that is, represented as photons, which can be created and destroyed) but the number of charged particles of a given sign is conserved. He inferred the form of the interaction between electrons and photons by going to the limit of large numbers of photons being present, in which case the interaction is essentially classical. In this respect, he used the idea of the Bohr Correspondence Principle, although it is not explicitly named.

Only in the fourth edition of 1958, but not earlier, did Dirac introduce a version of *QED* in which the number of charged particles is not fixed. This treatment used 'second quantization' of the electron-positron field, which allows the creation of these particles in pairs. In his preface, Dirac explained why he made this change:

> In present-day high-energy physics the creation and annihilation of charged particles is a frequent occurrence. A quantum electrodynamics which demands conservation of the number of charged particles is therefore out of touch with physical reality. So I have replaced it by a quantum electrodynamics which includes creation and annihilation of electron-positron pairs. This involves abandoning any close analogy with classical electron theory, but provides a closer description of nature.

Dirac remained skeptical about the validity of quantum field theory throughout his life, regarding *QED* and other local quantum field theories as mathematically unsound, as evidenced by their apparently infinite predictions for physically finite quantities such as mass and charge. In the opinion of most physicists, these defects were cured in the late 1940s by a relativistically covariant subtraction scheme known as renormalization. In the fourth edition of *Principles*, Dirac concluded his discussion of *QED* with the following comment (p. 309):

> People have succeeded in setting up certain rules that enable one to discard the infinities produced by the fluctuations in a self-consistent way and have thus obtained a workable theory from which one can calculate results that can be compared with experiment. Good agreement with experiment has been found, showing that there is some validity in the rules. But the rules are applicable only to special problems, usually collision problems, and do not fit in with the logical foundations of quantum mechanics. They should therefore not be considered a satisfactory solution of the difficulties.

Ten years after the fourth edition, in a '4th revised' edition, Dirac repeated that he did not see how the renormalization theory 'can be presented as a logical development of the standard principles of quantum mechanics'. He had found a way to calculate the Lamb shift and the anomalous magnetic moment of the electron (two successes of the renormalization theory) by a new formulation of *QED* that made use of the Heisenberg picture, in which the state vector is constant. The results agreed with the renormalization technique. Nevertheless, Dirac maintained that as regards high-energy physics a new approach would be needed, since 'we are effectively in the pre-Bohr era'.

5 WEYL, HILBERT, VON NEUMANN AND THE MATHEMATICAL FOUNDATIONS OF QUANTUM MECHANICS

Among the mathematicians interested in physical theories, Hermann Weyl (1885–1955), professor at the *Eidgenössische Technische Hochschule* (ETH) in Zurich, learned of the new matrix mechanics from Max Born in September 1925. He was impressed by what he called a 'fabulous discovery' and independently of Heisenberg (and others) proposed the commutation relations of matrices for several degrees of freedom. It is interesting to note that, although he had treated the problem of unbounded matrices having discrete plus continuous spectra in his doctoral thesis with Hilbert, he did not proceed to the equivalent differential-equation formalism or to the Schrödinger equation. Probably he was too deeply involved in his exhaustive program of finding the representations of all simple and semi-simple continuous groups. After he completed this task, he published some applications to quantum mechanics of these mathematical results in a paper of October 1927 [Weyl, 1927]. He showed there that 'the principal inner reason for the canonical pairing of quantum-mechanical variables shows up clearly in the case that the fundamental group [of the dynamics] is a continuous one, which still embraces discrete cases as well'. By emphasizing what he called 'the *integral standpoint* against the differential one' (that is, replacing the infinitesimal group by the full continuous group), he found that the tranformation from the equations of matrix mechanics 'to Schrödinger's wave equations can be performed with every necessary rigor'; he also removed the difficulties connected with the ordering of quantum-mechanical variables in products and thus established the kinematical structure of quantum mechanics on the basis of group theory.

Weyl further introduced the concept of *rays* in Hilbert space characterizing the pure states of physical problems, and discussed the classical groups in this quantum-mechanical 'ray space'. He summarized the results of this paper in a simplified manner in his well-known book *The theory of groups and quantum mechanics*, which emerged from a lecture course at the ETH [Weyl, 1928]. There he displayed the formalism of the rotation group, as well as that of the permutation group (including the Pauli principle), and presented Dirac's new relativistic theory of the electron. Although Weyl was not the first to introduce group-theoretical methods into quantum mechanics, having been preceded by Heisenberg and Eugene Wigner, it was mainly his book that helped the community of quantum physics to appreciate this mathematical theory as an important tool in solving atomic and high-energy physics problems.

A result of Weyl's group-theoretical approach, namely the solution of the general problem of canonical transformations in quantum mechanics and the demonstration of the equivalence of all existing different formulations of quantum mechanics, had already been considered in the fall of 1926 by Dirac (see section 1 above) and independently by Jordan. Jordan referred to Pauli's idea of taking the probability amplitude φ—which depends on two Hermitean variables—as the fundamental concept [Jordan, 1927]. He proposed a new statistical foundation of quantum mechanics resting on four postulates. On applying these postulates to the linear operators in quantum mechanics, he derived the laws of the theory in all existing schemes, from matrix mechanics to wave mechanics. In the last part of his lecture course in winter 1926–1927 Hilbert acknowledged Jordan's axiomatic approach in principle and claimed in addition: 'One does not need for the understanding of these

ideas a physical divination but just pure logic' [Hilbert, 1927, 204]. He first attacked this ambitious goal in a paper with two collaborators.

Early in 1926, Hilbert asked the International Education Board to grant a fellowship to a 22-year-old Hungarian mathematician, insisting that his candidate, Janos or John von Neumann (1903–1957), was a 'completely exceptional personality, who had already performed very productive work'. Born in Budapest on 28 December 1903, von Neumann had studied mathematics in Berlin (1921–1923) and afterwards chemical engineering at the Zurich ETH, receiving his Ph.D. (Budapest) and Master (Zurich) degrees in 1925; later he would apply for *Habilitation* in mathematics at Berlin University, submitting a thesis on 'The axiomatic construction of set theory' and passing it in December 1927. Von Neumann audited Hilbert's Göttingen lectures on the foundations of quantum theory, and wrote a joint paper with Hilbert and his physics assistant Lothar Nordheim on the mathematical foundations of quantum mechanics (1928). The authors stated their procedure as follows:

> One imposes certain physical conditions on the probabilities, which are suggested by our present knowledge. Then one seeks, in a second step, a simple analytic apparatus containing quantities that satisfy the same relations. This analytic apparatus, and therefore the quantities involved, now obtain on the basis of the physical requirements a physical interpretation. The goal is to formulate the physical requirements so completely that they fix the mathematical apparatus entirely.

They thus hoped to overcome the difficulties of the procedure, which the physicists had assumed, namely: guessing the mathematical apparatus of the theory before having recognized all physical requirements and then adjusting the mathematics. In contrast, the Göttingen trio demanded that one first determine the mathematical apparatus uniquely, and then apply a physical interpretation to the scheme.

In establishing a complete system of axioms for quantum mechanics, Hilbert and others closely followed Jordan's proposal, but they expanded his four axioms to six. Thus their Axiom I defines the *probability amplitude* $\Phi(x, y; F_1, F_2)$ in the case of two dynamical variables F_1 and F_2 , denoting that for a given value y of F_2, F_1 assumes a value between x and $x + dx$. Axiom II states that the square of this amplitude, the *probability density* $w(x, y; F_1, F_2)$, is identical to the one for the case, in which the value x of F_1 is given and F_2 assumes a value between y and $y + dy$. Axiom III demands that in the relation of a quantity F to itself a sharp determination is possible, that is, with each dynamical quantity a definite numerical value can be associated. Axiom IV establishes the addition and multiplication laws for probability amplitudes; Axiom V ensures that the probabilities are determined only by their functional dependence of the variables F_1 and F_2 on the canonical variable pair momentum p and position q, but not on the special properties of the Hamiltonian of the system; and Axiom VI requires that the probabilities for a given physical system do not depend on the coordinate system chosen.

Upon introducing a system of 'complete operators', which could be represented by integral operators with kernels, the authors defined two special operators $F(^{x}{}_{x})$ and $G(^{x}{}_{x})$,

which satisfied the functional equations

$$\left\{F\left(\begin{smallmatrix}x\\ x\end{smallmatrix}\right) - y\right\}\Phi(x, y) = 0 \quad \text{and} \quad \left\{G\left(\begin{smallmatrix}x\\ x\end{smallmatrix}\right) + \varepsilon \,\mathrm{I}\, \partial/\partial y\right\} = 0 \tag{8}$$

(with $\varepsilon = h/2\pi$, and I denoting the unit operator), and whose anticommutator was given by

$$GF - FG = \varepsilon \,\mathrm{I}. \tag{9}$$

The physical interpretation of the mathematical formalism was then obvious. From (8) the authors derived the time-independent Schrödinger equation, and from (9) the time-dependent Schrödinger equation. The solution of the time-independent Schrödinger equation, namely

$$\left\{H(\varepsilon\partial/\partial x, x) - W\right\}\Phi(x, W; q, H) = 0, \tag{10}$$

they wrote as

$$\Phi(x, y; q, H) = \sum_n c_n \psi_n(x)\delta(W - W_n) + \psi(x, W)c(W). \tag{11}$$

That is, the energy spectrum of the system described by the Hamiltonian H contains a sum of discrete terms and a continuous term, with c_n and $c(W)$ denoting the respective *a priori* probability amplitudes. 'Formally one can, by suitable application of the improper function, write the calculus as a matrix calculus and thus express clearly the genuine discrete nature of quantum mechanics, as Dirac has essentially done', the authors concluded but added that 'from the mathematical point of view the method of calculation used this way must be considered unsatisfactory, since one never knows to what extent the operations occurring can be really performed'.

In their last footnote, Hilbert et alii referred to a different approach, which von Neumann presented in a paper to the meeting of 20 May 1927 of the Göttingen Academy, entitled 'Mathematical foundations of quantum mechanics' [von Neumann, 1927a]. He justified his 57-page memoir by noting the weak points of the existing mathematical formalism of quantum mechanics. On the one hand, the matrix method did not really succeed in obtaining the eigenvalue spectrum of the quantum-theoretical Hamiltonian unless the spectrum was discrete: Dirac's introduction of continuous matrices could not be justified with mathematical rigor, because 'one must introduce with it concepts like infinitely large matrix elements or infinitely close diagonals'. On the other hand, Schrödinger's wave-mechanical scheme had to be complemented by the probability interpretation derived from matrix mechanics. Finally, both matrix and wave mechanics 'worked with quantities that were unobservable, hence senseless'. Even if the final results exhibited no such unsatisfactory features, they should be avoided, hence von Neumann announced a new 'method, which remedies these abuses and summarizes, in a unified and systematic way, the statistical point of view in quantum mechanics'.

To achieve this goal, the author exploited the available mathematical theory of Hilbert spaces. Unlike the physicists, he did not rely on the on the analogy between the discrete space of index values Z and the continuous state space Ω of a quantum-mechanical system, but identified Z with a complex Hilbert Space $\mathcal{H}_Z$ of denumerably many dimensions, and

Ω with another Hilbert space $\mathcal{H}_\Omega$ of square-integrable Schrödinger functions. However, since both have formally many properties in common, he put those into a third, the 'abstract Hilbert space $\mathcal{H}$', which is defined by the following axioms:

i) it is a linear space;

ii) it is a metric space, the metric arising from a symmetrical, bilinear form;

iii) it possesses infinitely many dimensions;

iv) there exists a sequence that is everywhere dense in $\mathcal{H}$ (later $\mathcal{H}$ would be called therefore a 'separable Hilbert space'); and

v) in $\mathcal{H}$ the Cauchy convergence condition applies. As an immediate consequence of these axioms, there followed the existence of a complete orthogonal system of functions φ.

In the abstract Hilbert space $\mathcal{H}$ von Neumann then introduced operators, especially what he called 'Einzel operators' (later renamed *projection operators*) $\underline{E}$ (with $\underline{E}^2 = \underline{E}$), which had already played a role in the eigenvalue problem, in the theory of linear integral equations. While Hilbert had shown in 1906 that every bounded symmetrical bilinear form possessed a *unique representation* with a *unique set of eigenvalues*, this was not yet clear for the unbounded symmetrical (Hermitean) operators occurring in von Neumann's space $\mathcal{H}$ for quantum mechanics. For the moment, he assumed this requirement to be valid and proceeded to demonstrate the 'mathematically rigorous unification of statistical quantum mechanics', that is, to establish the probability *Ansatz* used by Jordan in deriving the quantum mechanical transformation theory. Half a year later he removed these difficulties in his *Habilitation* thesis of December 1927 [von Neumann, 1929a]. Also the unbounded symmetrical operators could be shown, by limiting their range of definition in $\mathcal{H}$, to exhibit a unique eigenvalue spectrum. In a further paper he pointed to another necessary, yet weaker, restriction of the operators for that purpose: their matrices had to be 'square integrable', that is, the sums of the absolute squares of their row elements had to assume finite values.

By and large the rich and rigorous work performed by the young mathematical genius in 1927 substantiated the hopes of the theoretical physicists and satisfied the requirements of his mathematical colleagues. However, von Neumann emphasized in his thesis that 'the general theory of Hilbert operators does not reproduce everywhere the behavior generally expected and assumed in the "transformation theory" on the basis of the analogy with bounded or even finite dimensional operators' [von Neumann, 1929a, 62]. One might add that another Hungarian mathematician of his age, Aurel Wintner, attacked the same problem at the same time and arrived, between 1927 and 1928, at results consistent with those of von Neumann [Wintner, 1929].

Apart from the thorough discussion of the peculiar Hilbert space problems in quantum mechanics, von Neumann studied another path to a different axiomatic foundation of quantum mechanics. In a memoir on 'Probability-theoretical formulation of quantum mechanics', presented in November 1927 to the Göttingen Academy [von Neumann, 1927b], he turned to a systematic derivation of the physical theory from the facts of experience, which were all connected with the probability description. That is, in contrast to the physicists, who had taken a *deductive* path of identifying the absolute square of the Schrödinger wave

function with the probability of the state described, and verified the agreement with experiment *a posteriori*, he chose the *inductive* path by assuming the unrestricted validity of the usual probability calculus—as followed from the experiment—and derived the quantum-mechanical relations. For establishing the proper basis for this enterprise, he introduced a number of useful definitions and wrote down the necessary postulates. The definitions were stimulated by concepts from classical statistical mechanics, such as 'uniform' or 'elementary disordered ensembles', 'knowledge of a system', and 'expectation values' for its properties. As a particularly valuable concept for describing the knowledge of a quantum-mechanical system (or of ensembles of them) he introduced a *Hermitean operator* U in terms of a *density matrix* and stated: 'All possible knowledge corresponds uniquely in a one-to-one relation to the definite, linear operator U. This correspondence is described by the expectation value formula

$$E(S) = \sum s_{\mu\nu}\bar{u}_{\mu\nu}, \tag{12}$$

where S and U are represented by the matrices $\{s_{\mu\nu}\}$ and $\{u_{\mu\nu}\}$, respectively' [von Neumann, 1927b, 225]. Lev Landau arrived at a similar density operator in a special case in 1927.

Von Neumann concluded 'that quantum mechanics is not only compatible with the usual probability calculus; but, with a few plausible assumptions added, it is even the only possible solution', and he listed 'these basic assumptions':

1. Every measurement changes the observed object, and two measurements always disturb each other, if they cannot be replaced by a single one.
2. However, the change created by a measurement is of such nature that this measurement always remains valid, that is, when it is repeated *immediately* afterwards, the same result will be obtained.
3. The physical quantities have to be described by functional operators satisfying simple formal rules.

With these preparations he proceeded to develop the thermodynamics of quantum-statistical ensembles [von Neumann, 1927c] and later treated—after previous pioneering work by the physicists Schrödinger and Pauli—the ergodic hypothesis and the so-called H-theorem in quantum mechanics [1929b]. All these results entered, in a final formulation, his great conclusive book on the mathematical foundations of quantum mechanics.

6 THE *MATHEMATICAL FOUNDATIONS OF QUANTUM MECHANICS*

Physicists reacted only moderately to the contributions of their mathematical colleague von Neumann. Wolfgang Pauli cited a few results in his handbook article on wave mechanics, and Pascual Jordan wrote in a review of Wintner's book: 'The first successful advances in the field of the [mathematical] problems we owe to him, on the one hand, and to J. von Neumann, on the other'. Some active quantum theorists gave greater attention to von Neumann's treatment of probabilities. Thus, Dirac employed the density matrix in several nonrelativistic and relativistic problems, while Schrödinger and Pauli were interested in his results on the H-theorem. Born and Jordan paid particular tribute to most of

von Neumann's mathematical foundation and translated a large part of his statistical arguments on the quantum-mechanical systems and their eigenvalues from the abstract operator language into the more familiar language (for physicists) of vectors and tensors [Born and Jordan, 1930, esp. ch. 6]. However, to reach a larger scientific public, von Neumann wrote his own book on mathematical foundations, outlining the main motivation in the preface (1932, English translation, p. ix):

> Dirac, in several papers, as well as in his recently published book, has given a representation of quantum mechanics which is scarcely to be surpassed in brevity and elegance, and which is at the same time of invariant character. [. . .] The method of Dirac (and this is overlooked today in a great part of the quantum mechanical literature, because of the clarity and elegance of the theory) in no way satisfies the requirements of mathematical rigor—not even if these are reduced in a natural fashion to the extent common elsewhere in theoretical physics.

Von Neumann then criticized especially the introduction of the improper delta-function for the task of diagonalizing all self-adjoint operators in quantum mechanics; he insisted 'that the correct structure need not consist in a mathematical refinement and explanation of the Dirac method, but rather that it requires a procedure differing from the very beginning, namely, the reliance on the Hilbert theory of operators'. That is, his *Mathematical foundations* should establish a mathematical approach to quantum mechanics quite different from Dirac's *Principles*. The book contains six chapters, summarised in Table 2.

In the three chapters following the short introduction in Chapter I—where the usual quantum-mechanical scheme of the physicists is outlined, including the transformation theory—von Neumann displayed in detail his mathematical foundation, worked out between 1927 and 1929. Chapters V and VI the author devoted primarily to the discussion of the physical interpretation of quantum mechanics and the process of measurement, which emerged from Heisenberg's discovery of the uncertainty relations and the ensuing philosophical considerations leading to Bohr's *principle of complementarity*. In their book, Born and Jordan [1930] had derived the measurement process from the matrix-mechanical formalism and concluded: 'It is the measurement which forces the atom to decide for itself

Table 2. Contents by chapters of von Neumann's book.
The pages of the German/English versions are given;
the English edition is cited.

Chapter	**Page**	**Title**
I	4/3	Introductory considerations.
II	18/34	Abstract Hilbert space.
III	101/196	Quantum statistics.
IV	157/295	Deductive development of the theory.
V	184/347	General considerations.
VI	222/417	The measuring process. [End 262/455.]

an eigenvalue of [a variable] B; on the other hand, A [a variable not commuting with B] then loses its previously defined value. One therefore notices that the analysis of the laws of microphysics leads necessarily to radical consequences. Even the conventional concepts of subject, object and reality lose their usual meaning' (pp. 324–325). Two years later, von Neumann gave a more thorough analysis of the measurement process, casting the Heisenberg and Bohr's Copenhagen interpretation into the language of his mathematical foundations.

Von Neumann prepared the discussion in Chapter V by considering the relation between the quantum-mechanical measurement process and irreversible phenomena in classical thermodynamics. In Chapter VI he started, like Born and Jordan before, from a calculation of the act of measuring a quantity in standard quantum mechanics and derived this result immediately: The measurement procedure turns a pure state—which when unobserved evolves according to the Schrödinger equation in a causal manner—into a mixture, that is, an irreversible change occurs. In agreement with an idea of Bohr, he ascribed this change to a *psycho-physical parallelism*, and von Neumann stated in his book (pp. 418–420, abbreviated):

> Measurement or the related process of subjective perception is a new entity relative to the physical environment and no reducible to the latter. Nevertheless, it is a fundamental requirement of the so-called principle of pyscho-physical parallelism that it must be possible to describe the extra-physical process of the subjective perception as if it were reality in the physical world—i.e. to assign to its parts equivalent physical processes in the objective environment, in ordinary space. That is, we must always divide the world into two parts, the one being the observed system, the other the observer. The boundary between the two is arbitrary to a large extent. It can be pushed arbitrarily deeply into the interior of the body of the actual observer.

In order to express these considerations properly, von Neumann divided the world (that is, the system considered in it) into three parts, part I denoting the observed system, part II the measuring instrument, and part III the actual observer. The quantum-mechanical evaluation of the measurement process then yielded the same result, independently of whether he treated first the combined system (I + II) and then system III, or first the system I and then the combined system (II + III)—that is, he proved that in the measurement process one could without any consequences shift the boundary (later called *von Neumann cut*) between the observed system, the measuring instrument and the observer. Further he concluded: 'The noncausal measurement process [involving essentially his unitary U operator in (12)] is not produced by any incomplete knowledge of the state of the observer' (p. 439).

Von Neumann's evaluation of the measurement process was accepted as *the* standard treatment in the following two decades. But after 1952 his procedure received new critical attention, and different descriptions, either within the standard quantum-mechanical formalism or violating it, have been proposed [Mehra and Rechenberg, 2001; Redai and Stöltzner, 2001]. Some physicists especially attacked a result of Chapter IV, namely the conclusion that 'hidden variables', introduced in order to restore classical causality in atomic theory, must be excluded. The author proved their impossibility by applying two

laws obeyed by quantum mechanical operators and his equation (12); and thus he demonstrated that extra variables, not contained in the Schrödinger equation, would change experimentally substantiated results. The new hidden-variable discussion, started by David Bohm in 1952, led to quantitative formulation in the *Bell inequalities* [Bell, 1964], but experiments clearly contradict them. Obviously, nature agrees with quantum mechanics as the proper description of atomic processes, and preserves the important contributions contained in von Neumann's *Mathematical foundations*.

BIBLIOGRAPHY

Bell, J. 1964. 'On the Einstein Podolsky Rosen paradox', *Physics, 1*, 195–200.

Born, M. and Jordan, P. 1925. 'Zur Quantenmechanik', *Zeitschrift für Physik, 34*, 858–888.

Born, M. and Jordan, P. 1930. *Elementare Quantenmechanik*, Berlin: Springer.

Born, M., Heisenberg, W., and Jordan, P. 1926. 'Zur Quantenmechanik. II', *Zeitschrift für Physik, 35*, 557–615.

Dirac, P.A.M. 1925. 'Fundamental equations of quantum mechanics', *Proceedings of the Royal Society of London, A109*, 642–653.

Dirac, P.A.M. 1926a. 'Quantum mechanics and a preliminary investigation of the hydrogen atom', *Ibidem, A110*, 561–579.

Dirac, P.A.M. 1926b. 'On the theory of quantum mechanics', *Ibidem, A112*, 661–677.

Dirac, P.A.M. 1927a. 'The physical interpretation of quantum dynamics', *Ibidem, A113*, 621–641.

Dirac, P.A.M. 1927b. 'The quantum theory of emission and absorption of radiation', *Ibidem, A114*, 243–265.

Dirac, P.A.M. 1928. 'The quantum theory of the electron. I.', *Ibidem, A117*, 610–624.

Dirac, P.A.M. 1995. *Collected works 1924–1948* (ed. R.H. Dalitz), Cambridge: Cambridge University Press. [Contains the preceding papers.]

Heisenberg, W. 1925. 'Über die quantentheoretische Umdeutung kinematischer und mechanischer Beziehungen', *Zeitschrift für Physik, 33*, 879–893.

Heisenberg, W. 1985. *Collected works*, vol. A1, Berlin: Springer.

Hilbert, D. 1912. *Grundzüge einer allgemeinen Theorie der linearen Integralgleichungen*, Leipzig: Teubner.

Hilbert, D. 1927. 'Mathematische Grundlagen der Quantentheorie', notes of the lecture course in winter 1926–1927 at Göttingen, Niedersächsische Staat- und Universitätsbibliotek, Göttingen.

Hilbert, D., von Neumann, J., and Nordheim, L. 1928. 'Über die Grundlagen der Quantenmechanik', *Mathematische Annalen, 98*, 1–30.

Jordan, P. 1927. 'Über eine neue Begründung der Quantenmechanik', *Zeitschrift für Physik, 40*, 809–838.

Kragh, H. 1990. *Dirac: A scientific biography*, Cambridge and New York: Cambridge University Press.

Mehra, J. and Rechenberg, H. 1982, 2000, 2001. *The historical development of quantum theory*, vols. 4, 6/1, 6/2, New York: Springer.

Redai, M. and Stöltzner, M. (eds.) 2001. *John von Neumann and the foundations of quantum physics*, Dordrecht: Kluwer.

Schrödinger, E. 1926a. 'Quantisierung als Eigenwertproblem. (Erste Mitteilung)', *Annalen der Physik, 79*, 361–376.

Schrödinger, E. 1926b. 'Über das Verhältnis der Heisenberg–Born–Jordansche Quantenmechanik zu der meinen', *Ibidem, 79*, 734–756.

Schrödinger, E. 1926c. 'Quantisierung als Eigenwertproblem. (Dritte Mitteilung)', *Ibidem, 80,* 437–490.

Sommerfeld, A. 1919. *Atombau und Spektrallinien,* Braunschweig: Vieweg. [2nd ed. 1921, 3rd 1922, 4th 1924, all revised. English trans. of 3rd edition: *Atomic structure and spectral lines*, London: Methuen, 1923.]

Van der Waerden, B.L. (ed.) 1967. *Sources of quantum mechanics*, Amsterdam: North-Holland.

Von Neumann, J. 1927a. 'Mathematische Begründung der Quantenmechanik', *Nachrichten der Gesellschaft der Wissenschaften zu Göttingen,* 1–57.

Von Neumann, J. 1927b. 'Wahrscheinlichkeitstheoretischer Aufbau der Quantenmechanik', *Ibidem,* 273–291.

Von Neumann, J. 1927c. 'Thermodynamik statistischer Gesamtheiten', *Ibidem,* 273–291.

Von Neumann, J. 1929a. 'Allgemeine Eigenwerttheorie Hermitischer Funktional-operatoren', *Mathematische Annalen, 102,* 49–131.

Von Neumann, J. 1929b. 'Beweis des Ergodensatzes und des H-Theorems in der neuen Mechanik', *Zeitschrift für Physik, 57,* 30–90.

Von Neumann, J. 1961. *Collected works,* 6 vols. (ed. A.H. Taub), Oxford: Pergamon. [Contains the preceding papers.]

Weyl, H. 1927. 'Gruppentheorie und Quantenmechanik', *Zeitschrift für Physik, 46***,** 1–46.

Weyl, H. 1928. *Gruppentheorie und Quantenmechanik,* 1st ed., Leipzig: Hirtzel. [2nd ed. 1931. English trans.: *The theory of groups and quantum mechanics* (trans. H.P. Robertson), London: Methuen, 1931.]

Wintner, A. 1929. *Spektraltheorie der unendlichen Matrizen,* Leipzig: Hirzel.

CHAPTER 70

B.L. VAN DER WAERDEN, *MODERNE ALGEBRA*, FIRST EDITION (1930–1931)

K.-H. Schlote

This was the first textbook of algebra systematically to present, on an abstract set-theoretic basis using axiomatic methods, the various areas of this subject as a homogeneous whole. Thereby it made a decisive contribution to the dissemination and fulfillment of new structural conceptions in algebra and in the whole of mathematics.

First publication. *Moderne Algebra. Unter Benutzung der Vorlesungen von E. Artin und E. Noether,* 2 volumes, Berlin: Verlag Julius Springer, 1930–1931. 243 + 216 pages.

Later editions. 2nd ed. vol. 1 1937, vol. 2 1940. 3rd ed. vol. 1 1951, vol. 2 1955. 4th ed. *Algebra,* vol. 1 1955, vol. 2 1959. Vol. 1: 6th ed. 1964, 7th ed. 1966, 8th ed. 1971, 9th ed. 1993. Vol. 2: 6th ed. 1993.

English translation. Modern algebras. In part a development from lectures by E. Artin and E. Noether (trans. from the 2nd ed. by Fred Blum, with revisions and additions by the author), New York: Ungar, 1949–1950. [2nd ed. 1953. Further edition: *Algebra* (trans. F. Blum and John R. Schulenberger), 1970.]

Japanese translation. Gendai daisûgaku. Ko gimbayashi, 3 vols., Tokyo: Shoko Shuppansha, 1959. [Further eds.: *Enshû gendai daisûgaku,* vol 1, *Ko Gimbayashi*; *Enshû gendai sûgaku,* vol. 2, *Ko Gimbayashi*. Tokyo: Tokyo tosho, 1967, 1971.]

Russian translation. Algebra (trans. A.A. Bel'skogo, ed. Ju.I. Merzlyakova), Moscow: Nauka, 1979.

Related articles: Dirichlet (§37), Cantor (§46), Weber (§53), Hilbert on number theory (§54), Dickson (§65).

1 BIOGRAPHICAL NOTES

While still a student, Bartel Leender van der Waerden (1903–1996) had the good fortune to meet and collaborate with two mathematicians whose work gave a new direction and

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

shape to algebraic research: Emmy Noether (1882–1935) and Emil Artin (1898–1962). He first studied mathematics and physics at the University of Amsterdam from 1919 to 1924, aiming to enter the teaching profession in accordance with the wishes of his father. He later named the algebraist Hendrik de Vries (1867–1954) as his most important teacher, although he also attended the lectures of the theorist on foundations Gerrit Mannoury (1867–1956), the invariant theorist Roland Weitzenböck (1885–1951), and Luitzen E.J. Brouwer (1881–1966), the famous topologist and founder of intuitionism. He continued his studies at Göttingen in 1924–1925. After his military service, he spent a semester at the University of Hamburg in 1926, and in the same year graduated under de Vries at Amsterdam. Following a lectureship at Göttingen in 1927 and a professorship at Groningen, he taught at the University of Leipzig from 1931 to 1945. After visiting positions at Baltimore and his birthplace Amsterdam, Zürich became the second important workplace of his life, where he taught and worked at the university from 1951 to 1972, succeeding Rudolf Fueter (1880–1950); later he became professor emeritus.

The marked influences on van der Waerden were above all those of de Vries in Amsterdam, Noether in Göttingen and Artin in Hamburg. de Vries taught him classical algebraic geometry, in particular the enumerative calculus of Hermann C.H. Schubert (1848–1911). He then learned the new abstract conception of algebra from Noether and Artin, and how the shape of important parts of algebra became suitably modified in this context. During his time at Göttingen and Hamburg, van der Waerden also came under the important influence of Hellmuth Kneser (1898–1973), Otto Schreier (1901–1929) and Erich Hecke (1887–1947). His main area of work was algebraic geometry, whose adaptation and precise foundation he made a self-appointed task, and to which he made fundamental contributions, from his doctoral thesis to the series of 20 articles ‘Zur algebraischen Geometrie’, which appeared between 1933 and 1971.

Appropriately, the book *Moderne Algebra* stands at the start of van der Waerden’s scientific publications. In it, as in later works, he demonstrated his ability to extract the case of a theory and describe it clearly. Examples are the monographs ‘Introduction to algebraic geometry’, ‘Group-theoretical methods in quantum mechanics’, and ‘Mathematical statistics’, in which he also presented important results of his own. Despite its success, van der Waerden was later to describe *Moderne Algebra* as ‘the wrong book’, in contrast to *Einführung in die algebraische Geometrie* [1939], which for him was ‘the right book’.

2 AIMS AND CONTENTS OF THE BOOK: FOUNDATIONS

van der Waerden described the purpose of the book in the following words (vol. 1, p. 1):

> The ‘abstract’, ‘formal’ or ‘axiomatic’ setting to which algebra owes its recently acquired renewal of impetus has led, above all in *field theory*, *ideal theory*, *group theory* and the *theory of hypercomplex numbers*, to insight into new relationships and far-reaching results. The main aim of the book is to introduce the reader to this whole circle of ideas. Thus, general concepts and methods occupy the foreground, while the individual results that form the substance of classical algebra fall into their proper places within the framework of the modern structure.

The book was divided into 17 chapters (see Table 1, which covers also the third and seventh editions). A major concern of van der Waerden was to arrange the material in such a way as to form a coherent picture in which the individual sections are perceived from the foundations more or less independently of each other. This led in a fairly natural way to a hierarchical arrangement of the individual chapters. After the introduction of a basic concept, there follows at the next level the deeper and more thorough development of the theory relating to this concept, as well as the study of further fundamental algebraic structures. Appended to this, at a third level, was the investigation of important special areas.

van der Waerden achieved this object in a masterful way. The first three chapters contained the basic concepts, roughly described by the headings sets, groups, rings, ideals and fields. In set theory, he forged a strictly abstract treatment of the necessary terminology and results, by adopting the standpoint of naïve set theory without becoming involved with a discussion of the fundamentals arising in this context. Naïve set theory here means the pragmatic approach chosen by many mathematicians whereby the various basic concepts are understood with a certain intuitive clarity but all elements leading to paradoxes are avoided (compare §46 on Cantor). As to foundations, van der Waerden contented himself with a reference to the third edition of *Einleitung in die Mengenlehre* by Abraham Fraenkel (1891–1965) and later to the joint work of David Hilbert (1862–1943) and Paul Bernays (1888–1977) on the foundations of mathematics (§77). The basic algebraic structures, such as groups, rings and fields, were then defined in terms of abstract sets. Between the elements of a set were defined one or two operations that must satisfy certain rules. For example, the definition of group began as follows (vol. 1, 15):

> A non-empty set G of elements of any kind (such as numbers, maps or transformations) is called a *group* if the following four conditions hold. 1. There is a *rule of combination* that assigns to any pair a, b of elements of G a third element of the same set, which is usually called the *product* of a and b and written ab or $a \cdot b$. (The product may depend on the order of the factors: ab is not necessarily equal to ba.)

The other three axioms involved properties of the product, namely, the fulfilment of the associative law $(ab)c = a(bc)$, the existence of at least one left identity e with $ea = a$ for all a in G, and for every a in G the existence of at least one left inverse a^{-1}, with $a^{-1}a = e$. A group was said to be *Abelian* if ab is always equal to ba, that is, the commutative law holds. van der Waerden thus combined the approaches of Walter Dyck (1856–1934) and Heinrich Weber (1842–1913) (§53) with the results of the American school in the study of axiom systems to produce a set-theoretically acceptable formulation. The definitions of ring, field and ideal followed a similar pattern, where in the case of the last-named the characterization as a subset of a ring was also used. While proceeding from an abstract basis, van der Waerden was nevertheless at pains to establish the relationship with traditional algebra and showed in practice how numerous known results appear in the abstract setting.

van der Waerden used simple theorems on polynomials with coefficients in an arbitrary commutative ring or field to introduce differential quotients of entire rational functions without recourse to considerations of continuity. Having derived certain interpolation formulae, he devoted himself in particular to the factorisation and irreducibility of polyno-

Table 1. Contents by chapters of van der Waerden's book.
An overview is provided of the variations in distribution of chapters between the first, third and seventh editions: the first changes to the second volume were made in the fifth edition. The first page of each chapter is indicated.

Ch.	**1st edition 1930–1931**	**3rd edition 1951–1955**	**7th edition 1966–1967**
1	Numbers and sets. 4	Numbers and sets. 3	Numbers and sets. 3
2	Groups. 15	Groups. 19	Groups. 13
3	Rings and fields. 36	Rings and fields. 41	Rings and fields. 33
4	Completely rational functions. 67	Completely rational functions. 73	Vector spaces and tensor spaces. 62
5	Field theory. 86	Field theory. 101	Entire rational functions. 84
6	Continuation of group theory. 132	Continuation of group theory. 146	Field theory. 110
7	Galois theory. 148	Galois theory. 163	Continuation of group theory. 146
8	Ordered and well-ordered sets. 192	Infinite field extensions. 191	Galois theory. 168
9	Infinite field extensions. 198	Real fields. 200	Ordered and well-ordered sets. 209
10	Real fields. 209–238	Evaluated fields. 248–295	Infinite field extensions. 215
11	Elimination theory. vol. 2, 1	Elimination theory. vol. 2, 1	Real fields. 234–263
12	General theory of ideals of commutative rings. 23	General theory of ideals of commutative rings. 18	Linear algebra. vol. 2, 1
13	Theory of polynomial ideals. 51	Theory of polynomial ideals. 46	Algebras. 33
14	Completely algebraic numbers. 86	Completely algebraic numbers. 74	Representation theory of groups and algebras. 78
15	Linear algebra. 109	Linear algebra. 98	General theory of ideals of commutative rings. 120
16	Theory of hypercomplex quantities. 149	Theory of hypercomplex quantities. 135	Theory of polynomial ideals. 155
17	Representation theory of groups and hypercomplex systems. 177–212	Representation theory of groups and hypercomplex systems. 167–219	Integral algebraic quantities. 175
18			Evaluated fields. 200
19			Algebraic functions of a variable. 234
20			Topological algebra. 266–292

mials. A central role was played by the proof, of which the basic idea goes back to Carl Friedrich Gauss (1777–1855), that if uniqueness of factorisation holds for an integral domain S, then it also holds for the polynomial ring $S[x]$. This part of the book concluded with a treatment of symmetric functions, that is, functions of the variables $x_1, \ldots, x_n$ that are invariant under any permutation of $x_1, \ldots, x_n$, and the representation of such functions in terms of the so-called 'elementary symmetric functions'.

3 THE CONSOLIDATION OF FIELD THEORY AND GROUP THEORY: GALOIS THEORY

In the chapters that follow, van der Waerden gave a comprehensive account of field theory and a continuation of group theory. In field theory he followed the classical model of Ernst Steinitz (1871–1928), in giving a survey of all possible types of fields and their relationships with one another along with fundamental results on the structure of fields, including the theory of Galois fields (finite commutative fields) and cyclotomic fields, and leading finally to the theorem on the primitive element. A key constituent here was the analysis of field extensions, that is, fields that could be constructed from a prescribed ground field by the adjunction of new elements. The ground field was usually assumed to be commutative. In describing the process of extension, set-theoretic terminology again turned out to be advantageous: the adjunction of an arbitrary set M is reduced to that of finite sets and the construction of a union of sets, namely, the union of all fields arising from the adjunction of finitely many elements of M. The adjunction of a finite set can be obtained by the successive adjunction of single elements. Thus it is possible to concentrate on the adjunction of a single element, which is known as a simple extension. The latter was further characterized as algebraic or transcendental according as there does or does not exist a (non-trivial) polynomial with coefficients in the ground field having the adjoined element as a zero.

Following some preliminary assertions on the linear independence of abstract quantities and the systems formed by them, which are important in linear algebra and in particular satisfy the Steinitz exchange condition, the properties of algebraic and transcendental extensions were further investigated. The theorem on the primitive element answered the question of when a finite commutative extension of a commutative ground field can be obtained by the adjunction of a single generating (primitive) element. It is worth pointing out that van der Waerden ended the chapter with a section on 'the execution of the field-theoretic operations in finitely many steps'. This may also be understood, independently of the undisputed mathematical motivation of these considerations, as a reaction to the discussions of foundations going on at the time (compare §71).

van der Waerden began the deeper study of group theory by extending the group concept to that of groups with operators. These are characterized by the fact that an ordinary group is considered together with a further set of abstract elements called 'operators', and a product is defined whereby an operator is combined with a group element: one says that the operator acts on the group element. The product must again be an element of the group and action of the operator is a homomorphism of the group into itself. He then develops the theory of normal series and composition series. A *normal series* of a group G is defined as a series: $G \supset G_1 \supset G_2 \supset \cdots \supset G_n = E$, where each G_i is normal in its predecessor

G_{i-1}. Such a series, in which it is not possible to insert between two adjacent terms another term differing from both, is called a 'composition series'. When every factor group G_{i-1}/G_i of two adjacent terms is abelian, the group G is said to be 'solvable'. As the most important assertions of this theory, van der Waerden proved the theorem, which went back to Schreier, on the existence of isomorphic refinements of any two normal series in an arbitrary group, and the Jordan–Hölder theorem on the isomorphism of any two composition series of a given group. Then from the definition of direct product of groups there follow some particular assertions about permutation groups that will be needed later in the exposition of Galois theory. The three chapters together comprised a treatment of abstract algebraic structures that will enable the later theory to be developed in greater generality and, in particular, all the necessary preparations were made for the construction of Galois theory on this abstract basis.

This theory occupied the next seven chapters, and inaugurated the treatment of the special themes according to the hierarchy mentioned earlier. The main result of the theory described the well-known correspondence between the finite separable extensions of a commutative ground field K and the subgroups of a finite Galois group. As van der Waerden later acknowledged [1972–1973, 246], his exposition of Galois theory followed exactly the presentation given by Artin in his lectures of 1926. He introduced the Galois group as the automorphism group of an extension field of K, thereby following the interpretation made by Richard Dedekind (1831–1916) at the end of the 1850s (compare §37). In particular, the proof of the main theorem required the use of the primitive element theorem, a step which displeased Artin and which he was able to avoid in his revision of the theory in 1938. Following the main theorem, van der Waerden elucidated more precisely the correspondence between subgroups and intermediate fields, and showed firstly how to obtain the intermediate field corresponding to a subgroup of the Galois group and vice versa, and secondly how the Galois group changes when the ground field is extended.

The theory of cyclotomic fields was then developed as an application of Galois theory. A *cyclotomic field* is a field generated by the nth roots ζ of unity, and the *cyclotomic polynomial* is the polynomial $\Phi_n(x)$ whose zeros are precisely the primitive nth roots of unity counted once each. $\Phi_n(x)$ is shown to be irreducible (over the field $\mathbf{Q}$ of rational numbers), and it was then a simple matter to determine the Galois group of $\mathbf{Q}(\zeta)$. van der Waerden went on to describe the intermediate fields in the case when n is a prime number, and showed how to construct the cyclotomic field by the successive adjunction of quantities obtained from the ground field. All of this is based on ideas developed by Gauss in Section 7 of the *Disquisitiones arithmeticae* to prove the solubility of the equation $x^n - 1 = 0$ (n prime) using a chain of subsidiary equations each of lowest degree (§22.5). Historically, these fundamental researches of Gauss exercised a strong influence on the development of algebra in general, and in particular on the study of solvability of equations by Niels Henrik Abel (1802–1829) (§29) and Evariste Galois (1811–1832). The exposition of cyclotomic theory also contains a proof that the regular 17-gon is constructible using ruler and compasses, a result which formed the starting point for the young Gauss in his study of cyclotomic theory.

Following a section on the treatment of the equation $x^n - a = 0$ ($a \neq 0$), van der Waerden used Galois theory to solve, in about ten pages, the problem that lay at the heart of algebra for many centuries: the solution of equations by radicals. No better illustration of

the fundamental change in algebra could be found. As well as proving the famous assertion of Abel that the general equation of degree n is not solvable by radicals for $n > 4$, he also presented the solution of equations of the second, third and fourth degrees on this abstract basis. If this were not enough, the old question of the constructibility of geometric objects by ruler and compasses, which was among the classical problems of antiquity, was also solved. Also in this context is a general analysis of the construction of regular polygons by ruler and compasses, in which the construction of the 17-gon mentioned above appears as a special case. A few remarks on the calculation of the Galois group brought this chapter to a close.

4 IDEALS AND HYPERCOMPLEX SYSTEMS: FURTHER BASIC ALGEBRAIC THEORY

It is appropriate at this point to diverge from the order of chapters in the book and say something about the remaining fundamental topics. These were general theory of ideals in commutative rings (ch. 12), linear algebra (ch. 15) and the theory of hypercomplex systems (ch. 16).

At the heart of ideal theory lay the classification of the divisibility property for ideals in commutative rings, in particular the answer to the question of whether the elementary divisibility properties of the ordinary arithmetic of integers carry over to more general rings. This theory was investigated and developed in abstract terms in the decade preceding the appearance of the book, principally by Noether, using the classical exposition of Dedekind as a starting point. This involved the study of such important properties of rings as the validity of the basis theorem or, which is equivalent, the ascending chain condition. The former asserted that every ideal in a ring has a finite basis, while the latter, in its original formulation, requires that every chain of ideals $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, \ldots$ in which each $\mathbf{a}_i$ is a proper divisor of $\mathbf{a}_{i-1}$, that is, $\mathbf{a}_i \subset \mathbf{a}_{i-1}$, had only finitely many terms. Alongside the concepts of greatest common divisor and lowest common multiple, the operations of product and quotient of ideals had also to be introduced. For example, the *product* of the two ideals $\mathbf{a}$ and $\mathbf{b}$ was defined to be the ideal generated by the products $a \cdot b$ ($a \in \mathbf{a}$, $b \in \mathbf{b}$). As might be expected, the familiar factorisation of an ideal as a product of powers of prime ideals does not carry over without change from the ring of integers to more general rings, and so it becomes necessary to find properties which characterize the simple constituents of ideals and lead to an analogous factorisation.

The role of such constituents is played by the primary ideals, which are defined by the property that when the product $a \cdot b$ belongs to the ideal $\mathbf{q}$ and a is not an element of $\mathbf{q}$, then there is a positive integer ρ such that $b^\rho \in \mathbf{q}$. Moreover, for every primary ideal there is a corresponding prime ideal. As a central result, van der Waerden finally proved that every ideal can be represented as the intersection of finitely many primary ideals. This representation can be chosen in such a way that it cannot be shortened and no two primary ideals corresponding to the same prime ideal can be combined to form a larger primary ideal. These primary ideals are called the ‘greatest primary components’ of the ideal. In this form the representation satisfies certain uniqueness conditions, in that the number of greatest primary components, and also the prime ideals corresponding to the components,

are uniquely determined. These considerations and with a study of divisibility in the case of rings with identity, especially in relation to coprime ideals, that is, ideals whose greatest common divisor is the whole ring, close the chapter.

The objects of study in linear algebra are linear forms, modules, vectors and matrices; van der Waerden presented the various well-known constituents of classical algebra in an abstract setting, in particular that of groups with operators. Moreover, the concept of linear independence, already encountered in field theory, along with the results associated with it, came naturally into play. Following some remarks on modules in general and the specialisation to modules of linear forms (that is, modules whose elements can be represented as linear combination of finitely many linearly independent elements), he treated the theory of vector spaces and linear mappings between them. Since these mappings can be described using matrices, this treatment also involved elements of linear algebra and led naturally to the theory of solutions of systems of linear equations. The domain of multipliers of the module, which corresponds to the domain of coefficients of the system of equations, is now assumed to be a field, and when the field is actually commutative, the theory of determinants can be used in the derivation of solubility criteria and formulae for the solutions. All these topics in linear algebra required only a brief description and a reference to the classical theory.

A more thorough treatment was accorded to the theory of elementary divisors and the proof of the basis theorem for finitely generated Abelian groups. The latter concerned the decomposition of such a group into a direct sum of cyclic groups. Here again, van der Waerden strove for the highest possible level of generality. For example, a proof was given of the elementary divisor theorem for non-commutative division rings. The chapter ended with the rudiments of the theory of representation and representation modules, and also the reduction to normal form of matrices and thus of quadratic and Hermitian forms over a commutative field. This included the treatment of the characteristic polynomial and characteristic equation of a matrix.

A ring which at the same time is a module of linear forms over a commutative field K, leads to the notion of hypercomplex systems, or algebras, which form another important class of algebraic objects. A key point in the theory of hypercomplex systems is the elucidation of their structure. To begin with, a hypercomplex system can be regarded as a group with operators, where two separate domains of operators must be distinguished: the domain K of multipliers of the module, and the hypercomplex system itself. To these there correspond different admissible subgroups, where in the second case the ideals of the ring comprise the admissible subgroups. As a preliminary to the study of their structure, the notion of ideal is then extended to algebras by defining admissible left, right and two-sided ideals with respect to the domain K of multipliers. Under these conditions the ideals of a hypercomplex system then satisfy the maximum and minimum conditions, that is, every non-empty set of ideals has a largest and smallest number.

This property plays a central role in the subsequent investigations. van der Waerden emphasized this in his analysis of structure, in that from a general starting point with arbitrary rings he ended up only with those that satisfy these maximum and minimum conditions. If a ring has a domain of operators (multipliers), then ideals are defined as admissible ideals. It should be mentioned that the maximum condition is equivalent to the ascending chain condition, which was earlier of great importance in the construction of ideal theory.

It turns out, however, that the minimum condition is considerably more restrictive than the maximum condition.

Two further important concepts were those of nilpotent ideals and the radical. The former are ideals of which some finite power is the zero ideal, and the latter is the set of elements of the ring that each generate a nilpotent two-sided ideal. This leads to the introduction of semi-simple rings as rings with a zero radical that satisfy the minimum condition for left ideals. The structure of semi-simple rings is completely described: they are rings with identity equal to a direct sum of simple left ideals. The simple left ideals are those that contain no proper admissible ideal other than the zero ideal, and in this sense are the smallest building blocks in the construction of the ring.

In this structure theory, van der Waerden generalized the remarks and results of Benjamin Peirce (1809–1880), Wilhelm Killing (1847–1923), Elie Cartan (1869–1951) and Joseph H. Maclagan Wedderburn (1882–1948). It is interesting that he based the elaboration of the theory in this section on that given by Noether in her lectures, and made no mention of the generalization of Wedderburn's theorem by Artin in 1927. The related investigation of the analogues decomposition into two-sided ideals yields the uniqueness of the ideals involved, and as a special case the theorem of Dedekind on the decomposition of a commutative ring with zero radical and satisfy the minimum condition into a direct sum of commutative fields that annihilate one another. The subsequent study of the structure of the ring of automorphisms of a completely reducible module or a completely reducible ring with identity and its characterization as a direct sum of matrix rings, brought to a close the structural analysis of semi-simple rings. A consideration of the effects on the properties of algebras under the formation of products and the extension of the ground field concluded the chapter and paved the way for the special studies.

5 SOME SPECIAL TOPICS IN ALGEBRA

The chapter on ordered and well-ordered sets occupied a special place. Although it takes up only six closely packed pages, it is none the less very remarkable, as it supplies further basic notions of set theory, such as the axiom of choice, the well-ordering principle and transfinite induction. van der Waerden did not baulk at using elements of transfinite set theory in order to achieve a high level of abstraction. However, the application of these set-theoretic methods was not uncontroversial. As to the axiom of choice, for instance, it is assumed on the basis of examples that, given an arbitrary set of non-empty sets, one can define a choice function, that is, a function assigning to each set one of its members.

Using these additional set-theoretic tools, it is possible to develop the theory of infinite extensions, with which the treatment of special topics continues after the Galois theory. The majority of these topics concerned a detailed study of field theory. For infinite algebraic extensions, van der Waerden derived the theorem on the existence of an algebraically closed algebraic extension field, which is uniquely determined up to equivalence of extensions. A field is said to be algebraically closed if every member of the corresponding polynomial ring splits into linear factors. An algebraically closed extension field of a field K contains all possible algebraic extensions of K up to equivalence. The classification of the extensions of K is then completed by showing that an arbitrary infinite extension splits

into a pure transcendental and an associated algebraic extension. The purely transcendental extension is the result of adjoining to the ground field a set of algebraically independent unknowns. van der Waerden proved the additivity of the degree of transcendence for such extensions, but only in the case of finite degree, since the case of infinite degree would require the definition of addition of infinite cardinals. Here he followed the key points in the classical exposition of the theory; it was given by Steinitz in his work of 1910, which was fundamental to the development of algebra.

To this was added the theory of real fields, which goes back to Artin and Schreier. Regarding ordered fields, fields with a valuation and the definition of the real numbers, van der Waerden studied the non-algebraic properties involved and was at pains to explain the algebraic aspects of this theory. The result was a characterization of the basis arithmetic properties of the field of real numbers in purely algebraic terms. A formally real field can be defined as an abstract field in which -1 cannot be written as a sum of squares. Such a field is said to be 'real closed' if no proper algebraic extension of it is formally real. For such fields, van der Waerden was then able to supply algebraic proofs of a number of properties and theorems that are used in analysis, such as the existence of a unique ordering, Rolle's theorem and the theorem of Sturm on the number of distinct zeros in an interval. He also described the connection between the properties of real closure and algebraic closure: a real-closed field is not algebraically closed, but becomes so after the adjunction of an imaginary unit. Furthermore, in an algebraically closed extension Ω of a formally real field there is at least one intermediate field P for which $\Omega = P(i)$. Taking the ordering into account, one obtains the following assertion, which is important in the construction of the number system: an ordered field K has, up to isomorphism, only one real-closed algebraic extension field P in which the ordering is an extension of the ordering on K. Moreover, the identity is the only automorphism of P fixing the elements of K. Taking K to be the field of rational numbers, one is thereby able to construct the field of real algebraic numbers in a purely algebraic way.

A central role in algebraic geometry is played by the study of the solvability of algebraic equations in several unknowns and the derivation of formulae for their solutions; they form another collection of topics. A similar role is taken by resultants on polynomials whose vanishing is in general necessary and sufficient for the existence of a non-trivial solution. Invoking Leopold Kronecker's method of elimination and using the system of resultants for a system of polynomials in one variable, van der Waerden established a criterion for the solvability of such a system of polynomials. Further criteria concerned the solubility of one or more polynomial equations in n unknowns, as well as some specific statements on this problem. An important result, which is at the same time a generalization of successive elimination, is Hilbert's 'Nullstellensatz' of 1893, which asserted that a polynomial g in n unknowns over a commutative field belongs to the ideal generated by k polynomials f_i when g vanishes at all common zeros of the f_i, where $1 \leqslant i \leqslant k$.

The clarification of this kind of membership for polynomials played an important role in setting up the theory of polynomial ideals, which was itself one of the historical roots of the construction of ideal theory. On the other hand, the theory of polynomial ideals at the level of abstraction now achieved was a combination of results obtained in the general theory of ideals and those of field theory; so it formed the starting point for a deeper study of algebraic geometry. Next to the basic concepts, such as algebraic function, alge-

braic variety as the zero set of an ideal in the polynomial ring over a commutative field, and irreducibility of a variety, there stood a generalization of Noether's fundamental theorem, its application and a first introduction to intersection-point problems for algebraic varieties. The irreducibility of a manifold turned out to be equivalent to the existence of a representation of parameters by algebraic functions, and also to the fact that the associated ideal is prime. Using a method going back to Emanuel Lasker (1868–1941), it was further shown that every algebraic variety can be represented uniquely as the union of finitely many irreducible varieties.

A second historical root of ideal theory was the study of algebraic integers in algebraic number theory. van der Waerden took this into account in Chapter 14, sketching important results of this approach. For this, he extended the ascending chain condition to chains of modules over a ring R, and proceeded to characterize, in the case when R is a subring of a larger ring S, those elements of S that were integral over R. If S is commutative, then the integers again form a ring and the transitive law for integralness holds. After studying ideal theory in the ring of integers, where some assumptions are necessarily made about the ring R, which is now assumed to be an integral domain, there followed the axiomatic characterization of classical ideal theory in the form described by Noether in the mid 1920s. Ideal theory was constructed on the basis of three axioms: the ascending chain condition for ideals, the maximality of prime ideals, and integral closure. The meaning of each axiom was described and the unique prime factorisation of ideals deduced. This last result had a converse: the validity of the three axioms follows from the unique prime factorisation.

The book ended with a treatment of the representation theory of groups and algebras, which followed on immediately from the sections on linear algebra and the theory of hypercomplex systems. The representation problem for groups is reduced to that for hypercomplex systems by forming the group ring associated with a group and proving that every representation of the group is determined by a representation of the group ring.

An important role was played here by questions of the reducibility of representations and the theorem of Heinrich Maschke (1853–1908) on the complete reducibility of the representations of a finite group, as well as by the theory of characters. With regard to reducibility, two general theorems yield the assertions that every representation of a completely reducible hypercomplex system with identity is itself completely reducible, and every irreducible representation of an arbitrary algebra occurs in the regular representation. Characters are defined as the traces of irreducible representations over an algebraically closed field, where 'trace' means the trace of the matrix representing a given element. The use of characters makes it easy to obtain many properties of representations, in particular the fact that, under certain conditions, completely reducible representation of an algebra is uniquely determined up to equivalence by the traces of the representing matrices. The representations of some finite groups, such as the quaternion group and the symmetric groups S_3 and S_n, were derived by way of example. In conclusion, there was an application due to Noether of representation theory to the theory of non-commutative fields.

6 RECEPTION AND HISTORICAL IMPACT OF THE BOOK

van der Waerden's book effected a fundamental change in algebra and revolutionized mathematicians' perception of algebraic problems. It opened up a 'new world', as he had him-

self experienced when he came into contact with the new ideas of Göttingen in 1924 [van der Waerden, 1975, 32]. Other mathematicians said that after the appearance of the book the mathematical world was different from what it had been before. Algebra and algebraic problems seemed suddenly to take up a central position in mathematical research, and were no longer regarded as problems of peripheral interest. Garrett Birkhoff (1911–1996) described this in the following words [Birkhoff, 1973, 771]:

> Even in 1929 its concepts and methods (i.e. of 'modern' algebra) were still considered to have marginal interest as compared with those of analysis in most universities, including Harvard. By exhibiting their mathematical and philosophical unity and by showing their power as developed by E. Noether and her other younger colleagues (most notably E. Artin, R. Brauer, and H. Hasse) van der Waerden made 'modern algebra' suddenly seem central in mathematics. It is not too much to say that the freshness and enthusiasm of his exposition electrified the mathematical world—especially mathematicians under 30 like myself.

Although it did not fit into the historical picture, the book was for many the symbol of both the progress in algebra in the 20th century and the increased penetration of algebraic ideas and methods in other areas of mathematics. It is certain that the book substantially accelerated this progress, although it was only one factor among many others.

In a nutshell, van der Waerden's *Moderne Algebra* made its impression less by describing new results than by systematically putting together algebraic knowledge gained in the preceding decades and the resulting consideration of the new abstract approach and the application of axiomatic methods. The book exhibited not just one algebraic theory in this light, but the entire family of newly formed theories. Its significance was simply that it threw into bold relief the changed interaction between the branches of algebra and the resulting new image of the subject. The theory of the general equation of degree n, which a century earlier had been the main topic in algebra, was now treated in a few pages as an application of the abstractly formulated Galois theory.

The new image of algebra as a family of theories working together evolved as a general concept of structure with far-reaching effects on the whole of mathematics and other branches of science. In particular, many of these ideas were taken up and further developed by the influential Bourbaki group. Important elements of the structure concept were emphasized in the layout of the book without defining the concept of algebraic structure. Through the application of axiomatic methods, one clearly recognizes the properties characterizing algebraic conception at that time. This makes it possible to concentrate afresh on the particular 'structural properties' of objects alluded to in algebraic investigations and to recognize easily the occurrence of the same basic algebraic structure in various applications.

The definition of the basic algebraic objects followed either as sets of abstract elements admitting operations with certain specified properties, or as derived objects constructed from a given object by a definite algebraic procedure: the field of quotients is an example of the latter. The structural point of view is supported by an appropriate definition of mappings between algebraic objects in the form of homomorphisms and isomorphisms, with the understanding that isomorphic objects have the same algebraic properties and

thus embody the same structure. It must be pointed out, however, that van der Waerden defined these mappings separately for the individual structures. Likewise, homomorphisms of groups and rings are distinguished in the index of terminology, as are module and operator homomorphisms, and the mappings for the object under consideration are described more than once in the text.

Another feature was the emphasis on the consistent algebraic compilation of the theory, almost without reference to non-algebraic elements. This is especially apparent in, for example, the application of particular number systems. When the rational or real numbers appeared in an algebra textbook prior to van der Waerden both domains were taken for granted. Now they were characterized algebraically as fields.

The success of *Moderne Algebra* comes down finally to its lively style of presentation. Because of the abstract basis, the presentation of the individual theories is shorter and clearer, and thus has a stimulating effect on many readers. This is shown particularly clearly in a comparison with other algebra books of that time. His book was by no means the only textbook on algebra currently extant. On the contrary, the great progress made in algebra since the turn of the century had created an ever-increasing need for a new, systematic and comprehensive treatment of the results and theories. The standard text in the preceding decades, Weber's three-volume *Lehrbuch der Algebra* of 1895–1908 (§53) was rendered more and more out of date by these developments. From the middle of the 1920s on, there appeared among others the following textbooks, also written by famous algebraists: *Modern algebraic theories* by Leonard Eugene Dickson (1874–1954), *Höhere Algebra* (two volumes) by Helmut Hasse (1898–1979), *Algebra* (two volumes) by Oskar Perron (1880–1975), and *Einführung in die Algebra* (two volumes) by Otto Haupt (1887–1988). The first two of these books were published in 1926 and Perron's book followed a year later, and both Haupt's book and the German translation of Dickson's algebraic theories appeared in 1929. But none of these authors succeeded in describing the substance of the new conception of algebra as clearly as van der Waerden, or in interweaving it so elegantly with the body of classical results. In all their efforts to describe clearly and systematically the progress in algebra, they maintained a more or less strong commitment to the classical theory, and this hampered their exposition of the material. We note in passing that the first textbook in English to present algebra in the style of van der Waerden was *A survey of modern algebra* by Birkhoff and Saunders Maclane (b. 1909), which was published in 1941.

7 FURTHER REVISION OF THE BOOK

In view of the great influence of *Moderne Algebra* on the development of the subject, it is not surprising that a new edition very quickly became necessary and the first translation was soon made. The book rapidly became the standard textbook of algebra. With this in mind, van der Waerden made a special effort to take into account the further developments in algebra for the first of the new editions, and to take into consideration all the innovations in the care material of algebra indispensable in an introductory survey.

The second revised edition of the first volume appeared in 1937, and of the second in 1940. van der Waerden made some significant changes in his revision of the text. In the

first place, he reorganized the first volume into ‘a useful textbook of algebra for beginners’. He thus incorporated into it Euler’s theory of resultants and the theory of linear equations from the second volume, and added a section on decomposition into partial fractions. The notions of vector space and hypercomplex systems were now likewise explained in the first volume. He also derived the basic theorems on dimension, norm and trace alike in full generality, and included some further supplements and revisions.

The more detailed and thorough treatment of valuation theory was due to an enhanced understanding: it now had a chapter to itself. van der Waerden also reacted to the continuing discussion on the foundations of mathematics and the application of transfinite proof-methods. He tried to avoid such applications of set theory to algebra, and left out all those parts of field theory that involve the axiom of choice or the well-ordering theorem. Thus he also dropped the chapter on ordered and well-ordered sets. He obviously saw this as a suitable compromise with which to counteract any criticism. A completely finitistic development of algebra without recourse to any non-constructive existence proofs would have required too great a sacrifice (vol. 1, 2nd ed. 1937, p. vi).

All the chapters in the second volume were also subjected to a thorough revision, and in the Foreword van der Waerden particularly emphasized the considerable extension and remodelling of the chapter on hypercomplex numbers and their representations. This changed perception caused the second volume to assume more strongly the character of a deeper presentation concentrating particular areas of research, and the changes made were chiefly contingent on advances in individual areas of algebra. From this basic reorganization of the book and the transformation of the first volume into an introductory text, there arose in the following decade the need for a new edition of the first volume more often than of the second.

van der Waerden also revised the text from time to time for later editions of the book. The alterations and supplements were, however, not as extensive as those in the second edition. This is a clear reflection of the fact that the reshaping of algebra with regard to the statement of its problems and the formation of its fundamental branches, including their relationship to one another and to other areas of mathematics, was so far advanced that a definite stock of concepts and results had emerged as the core of the theory and achieved a far-reaching consensus among mathematicians. In the third edition (1951) he returned to the standpoint of the first in regard to the application of set-theoretic methods and revoked the constraints made in the second.

Since the methods of abstract algebra had in the meantime become the common property of mathematicians, van der Waerden followed a suggestion of Heinrich Brandt (1886–1954) and called merely ‘Algebra’ the book that appeared in 1955 as the first volume of the fourth edition. The second volume of the third edition, which was printed in the same year, likewise carried the changed title. The fourth edition incorporated in the second volume (1959) the important innovations stemming pre-eminently from progress in algebraic geometry. In two newly added chapters, he dealt with algebraic functions of one variable and with topological algebra. He developed the theory of algebraic functions up to and including the Riemann-Roch theorem for arbitrary fields of constants. Topological algebra was originally devoted to the completion of topological groups, rings and skew fields, including locally bounded and locally compact skew fields, and oriented itself chiefly around the fundamental work of

David van Dantzig (1900–1959) in 1933. Further revisions were made in the chapters on general ideal theory, algebraic integers and representation theory, which, by the completion of important ideal-theoretic theorems of W. Krull and the highlighting of new aspects and improved proofs, were brought to the very frontiers of research.

The former chapter on hypercomplex numbers was considerably extended and now renamed 'Algebras', a designation which has become established for this part of algebraic research. van der Waerden presented the main features of the theory of radicals developed by Nathan Jacobson (1910–1999) in the 1940s, and was able to provide a simplified proof of the main theorem by continuing ideas of Noether with the new contributions of Jacobson. At the same time, Grassmann algebras and Clifford algebras also found a place, whereas the chapter on elimination theory was dropped. In the fifth edition, the chapters of the second volume were completely reordered into three relatively independent groups. Chapters 12–14 were now devoted to linear algebra, algebras and representation theory, chapters 15–17 dealt with ideal theory and chapters 18–20 covered fields, algebraic functions and topological algebra.

Through all these variations and completions, van der Waerden's algebra has always remained topical and, despite strong competition, is a standard work in the corner of academic literature. Those who wish to scale the heights of algebra and its multifaceted applications must first become acquainted with van der Waerden's *Algebra*.

BIBLIOGRAPHY

Birkhoff, G. and Maclane, S. 1941. *A survey of modern algebra,* New York: Macmillan.

Birkhoff, G. 1973. 'Current trends in algebra', *American mathematical monthly, 80,* 760–782.

Corry, L. 1996. *Modern algebra and the rise of mathematical structures*, Basel: Birkhäuser.

Dickson, L.E. 1926. *Modern algebraic theories,* Chicago and New York: Sanborn.

Dickson, L.E. 1929. *Höhere Algebra* (trans. Ewald Bodewig), Leipzig and Berlin: Teubner.

Dold-Samplonius, Y. 1997. 'In memoriam. Bartel Leendert van der Waerden (1903–1996)', *Historia mathematica, 24*, 125–130.

Fraenkel, A. 1919. *Einleitung in die Mengenlehre. Eine gemeinverständliche Einführung in das Reich der unendlichen Größen*, Berlin: Springer. [2nd ed. 1923, 3rd ed. 1928.]

Hasse, H. 1926. *Höhere Algebra*, 2 vols., Berlin: de Gruyter.

Haupt, O. 1929. *Einführung in die Algebra*, Leipzig: Akademische Verlags-Gesellschaft.

Hlawka, E. 1995–1996. 'Bartel Leendert van der Waerden', *Almanach Österreichische Akademie der Wissenschaften, 146,* 399–405.

Perron, O. 1927. *Algebra*, 2 vols., Berlin and Leipzig: de Gruyter.

Schappacher, N. 2003. 'Bartel van der Waerden's work on algebraic geometry', *Streaming video,* www.msri.org/publications/video/.

Scriba, C.J. 1997. 'Bartel Leendert van der Waerden (2 Februar 1903 – 12 Januar 1996)', *Berichte zur Wissenschaftsgeschichte, 19,* 245–251.

van der Waerden, B.L. 1932. *Die gruppentheoretische Methode in der Quantenmechanik*, Berlin: Springer.

van der Waerden, B.L. 1939. *Einführung in die algebraische Geometrie*, Berlin: Springer.

van der Waerden, B.L. 1957. *Mathematische Statistik*, Berlin, Göttingen and Heidelberg: Springer.

van der Waerden, B.L. 1972–1973. 'Die Galois-Theorie von Heinrich Weber bis Emil Artin', *Archive for the history of exact sciences, 9*, 240–248.

van der Waerden, B.L. 1975. ‘On the sources of my book “Moderne Algebra” ’, *Historia mathematica, 2*, 31–40.
van der Waerden, B.L. 1983. *Zur algebraischen Geometrie. Selected papers*, Berlin: Springer.
van der Waerden, B.L. 1997. ‘ “Meine Göttinger Lehrjahre” (mit einem Nachwort von Peter Roquette)’, *DMV- Mitteilungen*, no. 2, 20–27.
Weber, H. 1895–1908. *Lehrbuch der Algebra*, 3 vols., Braunschweig. Vieweg. [See §53.]

CHAPTER 71

KURT GÖDEL, PAPER ON THE INCOMPLETENESS THEOREMS (1931)

Richard Zach

Gödel's incompleteness results are two of the most fundamental and important contributions to logic and the foundations of mathematics. He showed that no axiomatizable formal system strong enough to capture elementary number theory can prove every true sentence in its language. This theorem is an important limiting result regarding the power of formal axiomatics, but has also been of immense importance in other areas, such as the theory of computability.

First publication. 'Über formal unentscheidbare Sätze der *Principia mathematica* und verwandter Systeme. I', *Monatshefte für Mathematik und Physik, 37* (1931), 173–198.

Manuscripts. Two early drafts in Gabelsberger shorthand, the typewritten manuscript, page proofs, galley and an offprint held in the Kurt Gödel Papers at Princeton University Library, New Jersey, USA.

Reprint. In Gödel, *Collected works*, vol. 1 (ed. S. Feferman and others), New York: Oxford University Press, 1986, 116–195 [opposite English translation 3)].

English translations. 1) By B. Meltzer in *On formally undecidable propositions of Principia Mathematica and related systems*, Edinburgh: Oliver and Boyd, 1962, 35–72. 2) By E. Mendelsohn in M. Davis (ed.), *The undecidable*, Hewlett, NY: Raven Press, 1965, 4–38. 3) By J. van Heijenoort in his (ed.), *From Frege to Gödel. A source book in mathematical logic*, Cambridge, MA: Harvard University Press, 1967, 592–617. [Approved by Gödel. Repr. with intro. by S.C. Kleene in Gödel, *Collected works*, vol. 1 (see above; also several related pieces there). Also in S.G. Shanker (ed.), *Gödel's theorem in focus*, London: Routledge, 1988, 17–47.]

Italian translations. 1) As 'Proposizioni formalmente indecidibili dei *Principia Mathematica* e di sistemi affini I', in E. Agazzi (ed.), *Introduzione ai problemi dell'assiomatica*, Milan: Vita è Pensiero, 1961, 203–228. 2) By E. Ballo and others in their (ed.), Gödel, *Opere (1929–1936)*, Turin: Bollati Boringhieri, 1999, 113–138.

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

Portuguese translation. As 'Acerca de proposições formalmente indecidíveis nos *Principia Mathematica* e sistemas relacionados', in M. Lourenço (ed.), *O teorema de Gödel e a hipótese do contínuo*, Lisbon: Fundação Calouste Gulbenkian, 1979, 245–290.

Spanish translations. 1) By M. Garrido and others as 'Sobre proposiciones formalmente indecidibles de los *Principia Mathematica* y sistemas afines', in *Cuadernos Teorema* 8, Valencia (Spain): Revista Teorema, 1980. 2) By J. Mosterín as 'Sobre sentencias formalmente indecidibles de *Principa Matematica* y sistemas afines', in his (ed.), Gödel, *Obras completas*, Madrid: Alianza, 1981, 45–90.

Japanese translation. As '*Principia Mathematica* ya sono kanrentaikei deno keisikiteki ni ketteihukanou na meidai nitsuite I', in K. Hirose and K. Yokota (eds.), *Gödel no sekai* [Gödel's world], Tokyo: Kaimei-sha, 1985, 165–202.

French translation. By J.B. Scherer as 'Sur les propositions formellement indécidables des Principia Mathematica et des systèmes apparentés I', in E. Nagel and J.R. Newman, *Le théorème de Gödel*, Paris: Editions du Seuil, 1989, 106–143.

Related articles: Peano and Dedekind on arithmetic (§47), Cantor (§46), Whitehead and Russell (§61), Hilbert and Bernays (§77).

1 GÖDEL'S LIFE AND WORK

Kurt Gödel was born on 28 April 1906 in Brünn, the capital of Moravia, then part of the Austro-Hungarian Empire; now Brno, Czech Republic. His father was a well-to-do part-owner of a textile company. Gödel attended the German *Gymnasium* in Brünn and in 1923 followed his elder brother to study at the University of Vienna. He first studied physics, but Philipp Furtwängler's lectures on number theory so impressed him that he switched to mathematics in 1926. His teachers quickly realized Gödel's talent, and upon the recommendation of Hans Hahn (1879–1934), who later became his supervisor, Gödel was invited to join the group of philosophers around Schlick which became known as the 'Vienna Circle'. Gödel regularly attended until 1928, and later remained in close contact with some members of the circle, especially Rudolf Carnap (1891–1970). His interest in logic and the foundations of mathematics was sparked around that time, mainly through Carnap's lectures on logic, two talks which L.E.J. Brouwer gave in Vienna in 1928, and Hilbert and Ackermann's *Grundzüge der theoretischen Logik* (1928).

One of the open problems posed by David Hilbert (1862–1943) in [Hilbert, 1929] was that of the completeness of the axioms of the 'engere Funktionenkalkül', the first-order predicate calculus. Gödel solved this problem in his dissertation, which was submitted to the University of Vienna in 1929 and appeared as [Gödel, 1930].

Gödel then set to work on the main open problem in Hilbert's foundational program, that of finding a finitary consistency proof for formalized mathematics. This led him to the discovery of his first incompleteness theorem. In September 1930, following a report on his dissertation work, he gave the first announcement of his new result in a discussion of the foundations of mathematics at the 'Tagung für Erkenntnislehre der exakten Wissenschaften' in Königsberg. John von Neumann, who was in the audience, immediately

recognized the significance that Gödel's result had for Hilbert's program. Shortly thereafter, he wrote to Gödel with a sketch of the second incompleteness theorem about the unprovability of the consistency of a system within that system; but by that time Gödel had also obtained this result and published an abstract of it. The second result showed that Hilbert's program could not be carried out, and gave a negative solution to the second problem in Hilbert's famous 1900 list of mathematical problems (§57): Gödel proved that there can be no finitary consistency proof for arithmetic. The full paper was submitted for publication on 17 November 1930 and appeared in January 1931, in a Vienna mathematics journal edited by Hahn. It was also accepted as Gödel's *Habilitationsschrift* in 1932, and he was made *Privatdozent* (unpaid lecturer) at the University of Vienna in 1933.

Throughout the 1930s, Gödel worked on topics in logic and the foundations of mathematics, and lectured often on his incompleteness results. In particular, he gave a course on these results during his first visit to Princeton during the academic year 1933–1934 which exerted a significant influence on the logicians there, especially Alonzo Church and his student Stephen C. Kleene. During the 1930s, Gödel settled a subcase of the decision problem for first-order logic (proving the decidability of the so-called 'Gödel–Kalmar–Schütte class'), showed that intuitionistic logic cannot be characterized by finitely many truth values (and in the process inventing the family of Gödel logics), gave an interpretation of classical arithmetic in intuitionistic arithmetic (thus showing the consistency of the former relative to the latter), and established some proof-theoretic speed-up results.

After the annexation of Austria by Nazi Germany in 1938, during a second visit to Princeton, the title of *Privatdozent* was abolished. Gödel's application for *Dozent neuer Ordnung* was delayed, and he was deemed fit for military duty. He and his wife Adele, whom he had married in 1938, obtained U.S. visas and emigrated in 1940. From that date on, Gödel held an appointment at the Institute for Advanced Study at Princeton. 1940 also saw the publication of his third major contribution to mathematical logic, the proof of the consistency of the axiom of choice and of the continuum hypothesis (compare §46 on Cantor with the other axioms of set theory). This work was also inspired by a problem set by Hilbert: The first in his 1900 list of problems had asked for a proof of Cantor's continuum hypothesis. Gödel's result, together with Paul Cohen's 1963 proof of the consistency of the negation of the axiom of choice and of the continuum hypothesis, gave a negative solution to Hilbert's first problem: the axioms of set theory do not decide the continuum hypothesis one way or the other.

From 1943 onward, Gödel became increasingly interested in philosophy and relativity theory. In 1944, he contributed a study of Russell's mathematical logic (compare §61) to the Russell volume in the series *Library of living philosophers*. In the 1950s, he published several contributions to general relativity theory; in 1958, his consistency proof of arithmetic by an interpretation using functionals of hereditarily finite type, the so-called '*Dialectica* interpretation', appeared in print. Much of his post-1940 work, however, remained unpublished, including his modal-logical proof of the existence of God.

In the last ten years of his life, Gödel was in poor health, both physical and mental. He suffered from depression and paranoia, to the point at which fear of being poisoned kept

him from eating. He died of 'malnutrition and inanition' in Princeton on 14 January 1978. For more on his life and work, see [Feferman, 1986] and [Dawson, 1997].

2 HILBERT'S PROGRAM, COMPLETENESS, AND INCOMPLETENESS

Gödel's ground-breaking results were obtained against the backdrop of the foundational debate of the 1920s. In 1921, reacting in part to calls for a 'revolution' in mathematics by the intuitionist Brouwer and his own student Hermann Weyl, Hilbert had proposed a program for a new foundation of mathematics. The program called for i) a formalization of all of mathematics in an axiomatic systems followed by ii) a demonstration that this formalization is consistent, that is, that no contradiction can be derived from the axioms of mathematics. Partial progress had been made by Wilhelm Ackermann and John von Neumann, and Hilbert in 1928 claimed that consistency proofs had been established for first-order number theory. Gödel's results would later show that this assessment was too optimistic; but he had himself set out with the aim of contributing to this program.

According to [Wang, 1987], Gödel attempted to give a consistency proof for analysis relative to arithmetic. For this, he needed a definition of the concept of truth in arithmetic to verify (in arithmetic itself) the truth of the axioms of analysis. But Gödel soon realized that the concept of truth for sentences of arithmetic cannot be defined in arithmetic. The was led to this result by considerations similar to the liar paradox, thus anticipating later work by Alfred Tarski. But *provability* of a sentence from the axioms of arithmetic *is* representable in arithmetic, and combining these two facts enabled Gödel to prove that every consistent axiomatic system in which provability was representable must contain true, but unprovable sentences. Gödel had apparently obtained this result in the summer of 1930. At the time, he represented symbols by numbers, and formulas and proofs by sequences of numbers. Sequences of numbers can be straightforwardly formalized in systems of type theory or set theory. At the occasion of the announcement of his incompleteness result in the discussion at Königsberg, von Neumann asked if it was possible to construct undecidable sentences in number theory. This suggested a possible simplification to Gödel, and indeed he subsequently succeeded in arithmetizing sequences by an ingenious use of the Chinese remainder theorem.

It had been assumed by Hilbert that first-order number theory is complete in the sense that any sentence in the language of number theory would be either provable from the axioms or refutable (that is, its negation would be provable); indeed, he asked for a proof of this in his lecture on problems in logic [Hilbert, 1929]. Gödel's first incompleteness theorem showed that this assumption was false: it states that there are sentences of number theory which are neither provable nor refutable. The first theorem is general in the sense that it applies to any axiomatic theory which is ω-consistent (defined in the next section), has an effective proof procedure, and is strong enough to represent basic arithmetic. The system for which Gödel proved his results is a version of the system of *Principia mathematica*. In this system, the lowest type of variables ranges over numbers, the usual defining axioms for successor, plus comprehension are available as axioms. However, practically all candidates for axiomatizations of mathematics, such as first-order Peano Arithmetic, the full system of *Principia mathematica*, and Zermelo–Fraenkel set theory satisfy these conditions, and hence are incomplete.

3 AN OUTLINE OF GÖDEL'S RESULTS

Gödel's paper is organized in four sections. Section 1 contains an introduction and an overview of the results to be proved. Section 2 contains all the important definitions and the statement and proof of the first incompleteness theorem. In Section 3, he discusses strengthenings of this result. Section 4 is devoted to a discussion of the second incompleteness theorem. For detailed treatments of these results, see [Smorynski, 1977] or [Hájek and Pudlák, 1993].

In Section 2, Gödel first sets up some necessary definitions, gives the axioms of the variant P of the system of *Principia mathematica* which he uses, introduces the machinery necessary for the arithmetization of metamathematics (Gödel numbering), and proves four theorems (I–IV) about recursive functions and relations. The language of the system P consists of the usual logical symbols, 0 and the successor function f, as well as a repository of simply typed variables. Variables of the lowest type range over natural numbers, variables of the next type range over classes of numbers, variables of the third type range over classes of classes of numbers, and so on. The axioms of the system are the usual logical axioms, the comprehension schema $(\exists u)(\forall v)(u(v) \equiv A(v))$ (where u is a variable of type $n+1$, v a variable of type n, and A a formula not containing u free), and the extensionality axiom $(\forall v)(x(v) \equiv y(v)) \to x = y$.

One of the novel methods that Gödel uses is the arithmetization of syntax, now called 'Gödel numbering'. In order to be able to formalize reasoning about formulas and proofs in system P—which is, after all, a system for number theory—Gödel defines a mapping of the symbols in the language of P to numbers. In his paper the mapping is given by $0 \mapsto 1$, $f \mapsto 3$, $\sim \mapsto 5$, $\vee \mapsto 7$, $\forall \mapsto 9$, $(\mapsto 11$, $) \mapsto 13$, and the kth variable of type n is mapped to p_k^n, where p_k is the kth prime > 13 (for example, the first variable of lowest type is coded by 17). A sequence of symbols (for example, a formula) with codes $n_1, \ldots, n_k$ is then mapped to the number $2^{n_1} \cdot 3^{n_2} \cdot \cdots \cdot p_k^{n_k}$.

What Gödel calls 'recursive functions and relations' would now be called *primitive* recursive functions (and relations); he used the terminology in use at the time. A function ϕ is primitive recursive if there is a sequence of functions each of which is either the successor function $x+1$, a constant function, or results from two functions ψ, μ occurring previously in the sequence by the schema of primitive recursion

$$\phi(0, x_2, \ldots, x_n) = \psi(x_2, \ldots, x_n), \tag{1}$$

$$\phi(k+1, x_2, \ldots, x_n) = \mu\big(k, \phi(k, x_2, \ldots, x_n), x_2, \ldots, x_n\big). \tag{2}$$

A relation between natural numbers is primitive recursive if it can be defined by $\phi(x_1, \ldots, x_n) = 0$, where ϕ is a primitive recursive function.

Gödel's Theorem I states that primitive recursive functions are closed under substitution and primitive recursion. Theorem II states that recursive relations are closed under complement and union. Theorem III states that if two functions ϕ, ψ are primitive recursive, then so is the relation defined by $\phi(\bar{x}) = \psi(\bar{x})$. Finally, Theorem IV establishes that primitive recursive relations are closed under bounded existential generalization, that is, if $\phi(x)$ and $R(x, \bar{y})$ are primitive recursive, then so is the relation defined by $(\exists z)(z \leqslant \phi(x) \& R(z, \bar{y}))$.

Gödel next defines 46 functions and relations needed for the arithmetization of syntax and provability, of which the first 45 are primitive recursive. These definitions culminate in the definition of (45) xBy ('x is a proof of y') and (46) $Bew(x)$ ('x is a provable formula'). $Bew(x)$ is not primitive recursive, since it is obtained from B by *un*bounded existential generalization (that is, as $(\exists y)yBx$). These definitions use the arithmetization of syntax introduced earlier in the sense that, for example, the relation $Bew(x)$ holds of a *number* x if it is the code of a provable formula.

Gödel then sketches the proof of Theorem V, which states that whenever R is a primitive recursive n-ary relation, then there is a formula A with n free variables, so that if $R(k_1, \ldots, k_n)$, then $A(\bar{k}_1, \ldots, \bar{k}_n)$ is provable, and when not $R(k_1, \ldots, k_n)$, then $\sim A(\bar{k}_1, \ldots, \bar{k}_n)$ is provable. (Here, $\bar{k}$ is 0 preceded by k f's.) A formula A that is obtained from the primitive recursive definition of R in the way outlined in the proof is called a 'primitive recursive formula'. Since the proof is only sketched, this is not an explicit definition of what a primitive recursive formula is. In particular, in the system P, the most natural way to formalize primitive recursion is by higher-order quantification over sequences of numbers. Gödel explicitly uses such a second-order quantifier in the proof of Theorem VII discussed below. The method of constructing a formula of P which satisfies the conditions of Theorem V for a given primitive recursive relation R—a formula which 'numeralwise represents' R—yields such a formula for each of the 46 functions and relations defined earlier.

Following the proof of Theorem V, Gödel introduces the notion of ω-consistency. Roughly, an axiomatic system is ω-consistent if one cannot prove both $A(\bar{n})$ for all n and $\sim(\forall x)A(x)$. Theorem VI then is the first incompleteness theorem. Suppose κ is a primitive recursive predicate that defines a set of (codes of) formulas, which we might add as axioms to system P. Then we can, in a similar way as before, define the relation $xB_\kappa y$ (x is a proof of y in P_κ) and the predicate $Bew_\kappa x$ (x is provable in P_κ). Theorem VI states that if P_κ is ω-consistent, then there is a primitive recursive formula $A(x)$ so that neither $(\forall x)A(x)$ nor $\sim(\forall x)A(x)$ are provable in P_κ. By an ingenious trick combining diagonalization and the arithmetization of syntax (especially Theorem V), Gödel proves that there is a formula $A(x)$ so that $(\forall x)A(x)$ is provably equivalent in P_κ to $\sim Bew_\kappa(\bar{p})$, where p is the Gödel number of $(\forall x)A(x)$ itself. Hence, $(\forall x)A(x)$ in a sense says of itself that it is unprovable.

The paper continues in Section 3 with a number of strengthenings of Theorem VI. The formula A whose existence was proved in Theorem VI may contain quantifiers over higher-type variables. A relation which can be defined using only quantification over individual variables, and also $+$ and $\cdot$ as additional functions (addition and multiplication) is called *arithmetical*. Theorem VII states that every primitive recursive relation is arithmetical. Furthermore, the equivalence of recursive relations with arithmetical relations is formalizable in the system, that is, if A is a primitive recursive formula, then system P proves that A is equivalent to an arithmetical formula (one containing $+$, $\cdot$, but no quantification over variables of higher type). It then follows from Theorem VI that every ω-consistent axiomatizable extension of P contains undecidable arithmetical sentences (Theorem X).

The final section is devoted to the second incompleteness theorem (Theorem XI), which says that the formalization of consistency of an extension P_κ of P is not provable in P_κ. In

this context, the formula formalizing consistency of P_κ is taken to be

$$Wid_\kappa \equiv (\exists x)\big(Form(x) \& \sim Bew_\kappa(x)\big) \tag{3}$$

('there is an unprovable formula'). The proof of Theorem XI is only sketched. The argument for the first half of Theorem VI, namely, that $(\forall x)A(x)$ is unprovable in P_κ, uses only the consistency of P_κ but not its ω-consistency. By formalizing this proof in P_κ itself, we see that P_κ proves the implication $Wid_\kappa \to \sim Bew_\kappa(\bar{p})$, where p is the Gödel number of the unprovable $(\forall x)A(x)$. But, as noted above, $\sim Bew_\kappa(\bar{p})$ is equivalent, in P_κ, to $(\forall x)A(x)$. So if were provable, then $(\forall x)A(x)$ would be provable as well.

4 IMPORTANCE AND IMPACT OF THE INCOMPLETENESS THEOREMS

The main results of Gödel's paper, the first (Theorem VI) and second (Theorem XI) incompleteness theorems, stand as two of the most important in the history of mathematical logic.

Their importance lies in their generality: although proved specifically for extensions of system P, the method Gödel used is applicable in a wide variety of circumstances. Any ω-consistent system for which Theorem V holds will also be incomplete in the sense of Theorem VI. Theorem XI applies not as generally, and Gödel only announced a second paper in which this was going to be carried out for systems which are not extensions of P. However, the validity of the result for other systems was soon widely recognized, and the announced paper was never written. Hilbert and Bernays [1939] provided the first detailed proof of the second incompleteness theorem, and gave some sufficient conditions on the provability predicate *Bew* in order for the theorem to hold (§77.4.2).

One important aspect of the undecidable sentence $(\forall x)A(x)$ is that, although it is neither provable nor refutable *in* P, it is nevertheless readily seen to be *true*. For what it states is that it itself is not provable in P, and by the first incompleteness theorem, this is precisely the case. Since it is also not refutable, that is, its negation is also unprovable in P, the existence of undecidable sentences like $(\forall x)A(x)$ shows the possibility of axiomatic systems which are ω-inconsistent. The system resulting from P by adding $\sim(\forall x)A(x)$ as an additional axiom is one example. It proves $A(\bar{n})$ for all n, and also $(\exists x)\sim A(x)$. Although by Theorem VI there will also be true but unprovable statements in *this* system, the existence of undecidable sentences is left open. Rosser [1936] weakened the assumptions of Theorem VI and showed that not only ω-inconsistent but also consistent systems of the type discussed by Gödel will contain independent sentences.

The immediate effect of Gödel's theorem, and in particular, of his second theorem, was that the assumptions of Hilbert's program were challenged. Hilbert assumed quite explicitly that arithmetic was complete in the sense that it would settle all questions that could be formulated in its language—it was an open problem he was confident could be given a positive solution. The second theorem, however, was more acutely problematic for Hilbert's program. As early as January 1931, in correspondence between Gödel, Bernays, and von Neumann, it became clear that the consistency proof developed by Ackermann must contain errors [Zach, 2003]. Both Bernays and von Neumann accepted that the reasoning in

Gödel's proof can be readily formalized in systems such as P; on the other hand, a consistency proof should, by Gödel's own methods, also be formalizable and yield a proof in P of the sentence expressing P's consistency.

The errors in the consistency proof were soon found. It fell to [Gentzen, 1936] to give a correct proof of consistency using methods that, of necessity, could not be formulated in the system proved consistent. Although Gödel's results dealt a decisive blow to Hilbert's program as originally conceived, they led to Gentzen's work, which opened up a wide range of possible investigations in proof theory. For more on the reception of Gödel's theorems, see [Dawson, 1989] and [Mancosu, 1999].

As mentioned above, up to 1930 it was widely assumed that arithmetic, analysis, and indeed set theory could be completely axiomatized, and that once the right axiomatizations were found, every sentence of the theory under consideration could be either proved or disproved in the object-language theory itself. Gödel's theorem showed that this was not so, and that once a sharp distinction between the object- and metatheory was drawn, one could always formulate statements which could be decided in the metatheory, but not in the object theory itself. The first incompleteness theorem shows that object-level provability is always outstripped by meta-level truth. Gödel's proof, by example as it were, also showed how carefully object- and meta-language have to be distinguished in metamathematical considerations. A few years later, Tarski's work on truth and semantic paradoxes pointed to the same issue, showing that truth cannot be defined in the object-level theory (provided the theory is strong enough).

Gödel's results had a profound influence on the further development of the foundations of mathematics. One was that it pointed the way to a reconceptualization of the view of axiomatic foundations. Whereas a prevalent assumption prior to Gödel—and not only in the Hilbert school—was that incompleteness was at best an aberrant phenomenon, the incompleteness theorem showed that it was, in fact, the norm. It now seemed that many of the open questions of foundations, such as the continuum problem, might be further examples of incompleteness. Indeed, in [Gödel, 1940] he succeeded not long after in showing that the axiom of choice and the continuum hypothesis are not refutable in Zermelo–Fraenkel set theory: Cohen [1966] later showed that they were also not provable. The incompleteness theorem also played an important role in the negative solution to the decision problem for first-order logic by Church [1936]. The incompleteness phenomenon not only applies to provability, but, via the representability of recursive functions in formal systems such as P, also to the notion of computability and its limits.

Perhaps more than any other recent result of mathematics, Gödel's theorems have ignited the imagination of non-mathematicians. They inspired Douglas Hofstadter's best-seller *Gödel, Escher, Bach* (1979), which compares phenomena of self-reference in mathematics, visual art, and music. They also figure prominently in the work of popular writers such as Rudy Rucker. Although they have sometimes been misused, as when self-described postmodern writers claim that the incompleteness theorems show that there are truths that can never be known, the theorems have also had an important influence on serious philosophy. John Lucas, in his paper 'Minds, machines, and Gödel' (1961) and more recently Roger Penrose in *Shadows of the mind* (1994) have given arguments against mechanism (the view that the mind is, or can be faithfully modeled by a digital computer) based on

Gödel's results. It has also been of great importance in the philosophy of mathematics: for instance, Gödel himself saw them as an argument for Platonism [Feferman, 1984].

BIBLIOGRAPHY

Church, A. 1936. 'A note on the Entscheidungsproblem', *Journal of symbolic logic, 1*, 40–41.

Cohen, P.J. 1966. *Set theory and the continuum hypothesis*, New York: W.A. Benjamin.

Dawson, J.W. 1989. 'The reception of Gödel's incompleteness theorems', in [Shanker, 1988], 74–95.

Dawson, J.W. 1997. *Logical dilemmas: the life and work of Kurt Gödel*, Wellesley, MA: A.K. Peters.

Feferman, S. 1984. 'Kurt Gödel: conviction and caution', *Philosophia naturalis, 21*, 546–562. [Repr. in his *In the light of logic*, Oxford: Oxford University Press, 1998, 150–164.]

Feferman, S. 1986. 'Gödel's life and work', in [Gödel, 1986], 1–36.

Gentzen, G. 1936. 'Die Widerspruchsfreiheit der reinen Zahlentheorie', *Mathematische Annalen, 112*, 493–565. [English trans. in his *The collected papers* (ed. M.E. Szabo), Amsterdam: North-Holland, 1969, 132–213.]

Gödel, K. 1930. 'Die Vollständigkeit der Axiome des logischen Funktionenkalküls', *Monatshefte für Mathematik und Physik, 37*, 349–360. [Repr. and English trans. in [Gödel, 1986], 102–123.]

Gödel, K. 1940. *The consistency of the axiom of choice and of the generalized continuum hypothesis*, Princeton: Princeton University Press (Annals of Mathematics Studies, vol. 3).

Gödel, K. 1986. *Collected works*, vol. 1 (ed. S. Feferman and others), New York: Oxford University Press.

Hájek, P. and Pudlák, P. 1993. *Metamathematics of first-order arithmetic*, Berlin: Springer.

Hilbert, D. 1929. 'Probleme der Grundlegung der Mathematik', *Mathematische Annalen, 102*, 1–9. [English trans. in [Mancosu, 1998], 266–273.]

Hilbert, D. and Ackermann, W. 1928. *Grundzüge der theoretischen Logik*, 1st ed., Berlin: Springer.

Hilbert, D. and Bernays, P. 1939. *Grundlagen der Mathematik*, vol. 2, Berlin: Springer. [See §77.]

Mancosu, P. (ed.) 1998. *From Brouwer to Hilbert. The debate on the foundations of mathematics in the 1920s*, Oxford: Oxford University Press.

Mancosu, P. 1999. 'Between Vienna and Berlin: The immediate reception of Gödel's incompleteness theorems', *History and philosophy of logic, 20*, 33–45.

Rosser, J.B. 1936. 'Extensions of some theorems of Gödel and Church', *Journal of symbolic logic, 1*, 87–91.

Shanker, S.G. (ed.) 1988. *Gödel's theorem in focus*, London: Routledge.

Smorynski, C. 1977. 'The incompleteness theorems', in J. Barwise (ed.), *Handbook of mathematical logic*, Amsterdam: North-Holland, 821–865.

Wang, H. 1987. *Reflections on Kurt Gödel*, Cambridge: MIT Press.

Zach, R. 2003. 'The practice of finitism. Epsilon calculus and consistency proofs in Hilbert's program', *Synthese, 137*, 211–259.

CHAPTER 72

WALTER ANDREW SHEWHART, *ECONOMIC CONTROL OF QUALITY OF MANUFACTURED PRODUCT* (1931)

Denis Bayart

In this book Shewhart brought together many of the statistical principles of industrial quality control. His work contributed largely to the opening up of the field of industrial applications to mathematics.

First publication. New York: Van Nostrand; London: Macmillan, 1931. xiv + 501 pages.

Photoreprint. '50th anniversary commemorative reissue', Milwaukee, Wisconsin: American Society for Quality Control, 1980.

Spanish translation. Control económico de la calidad de productos manufacturados, Madrid, Spain: Diaz de Santos, 1997.

Related articles: Laplace on probability (§24), Pearson (§56), Fisher (§67).

1 THE INDUSTRIAL PROBLEM OF THE STATISTICAL CONTROL OF QUALITY

This book is the first publication to deal specifically with the control of industrial manufacturing processes by means of mathematical statistics. It relies mainly on the theories of distributions and sampling. This introduction of mathematical analysis into the field of industrial operations gives rise to important developments in the engineering sciences and also leads to statistical decision theory.

The flourishing of mass production in industry during the 20th century has caused problems in controlling the quality of manufactured products. They are produced in numbers too great to allow individual control, and sometimes the control involves their destruction (for example, munitions). In other cases, the control requires extensive operations in the laboratory, which inhibits a quick reaction to correct the errors detected.

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

In this context, statistics and the calculus of probabilities furnish rational methods for controlling production by sampling. Probability makes it possible to calculate the risk of making a bad decision based on a sample of given size. These calculations have caused many surprises, as there is always a tendency to over-estimate how representative a sample is. Thus, to judge the quality of a batch it was customary to take a sample of 2% of the batch size, which is much too small to form the basis of a valid judgement. Moreover, Poisson's law has been rediscovered in the observation that the sample need not necessarily be proportional to the size of the batch.

Thus, in the course of the 1920s, many methods appeared independently in several industrial countries. The mathematics involved is not particularly new, and had indeed been known for a long time. Thus it was certainly the economic and industrial conditions that encouraged the development of these methods at this particular time.

It was W.A. Shewhart (1891–1967) who produced the most original and profound publication of his day. A physicist by training with a Ph.D. in physics from the University of California at Berkeley in 1917, he was recruited in 1918 by the Western Electric Company and then in 1925 by Bell Telephone Laboratories, where he remained until his retirement in 1956. As well as statistics, he was involved with standardization in liaison with several large engineering associations.

Shewhart also played an important part in the American statistical community: founding member (1937) and president (1944) of the Institute of Mathematical Statistics, and president (1945) of the American Statistical Association. In addition, he was the first editor of the series *Wiley publications in statistics*, where he placed emphasis on applied statistics. His works were also taken into the field of management by the statistician and consultant W.E. Deming (1900–1993): they enjoyed great success in Japan, where they contributed to the industrial miracle of the 1970s (quality circles, total quality control) before being re-imported to the United States in the 1980s.

2 AN APPROACH VIA STATISTICAL PHYSICS

The problem to be addressed is that of the variability of physical quantities brought into play in industrial processes. It is this variability that can explain the bad quality of products. Indeed, the good quality of a product is defined as its conformity to a set of specifications and tolerances (in this approach, the quality of industrial products is always assumed to be measurable). The manufacturing equipment, raw materials and manual operations introduce into the process a variability that is not measured, and may be incompatible with the tolerances imposed.

Shewhart studied this variability by regarding it as a property of the system of production and by representing it as a statistical distribution whose parameters could be estimated. But for such a hypothesis to be valid, it is necessary that a system of production be in a stable state, which Shewhart conceptualised under the name of 'constant system of chance causes'. In general, the production equipment in a factory is not in this state. To get around this, it is necessary to identify the 'assignable causes of variation' and by eliminating them to reduce the variability and regularize production. He referred explicitly to P.S. Laplace in these first reflections [Shewhart, 1924, 57].

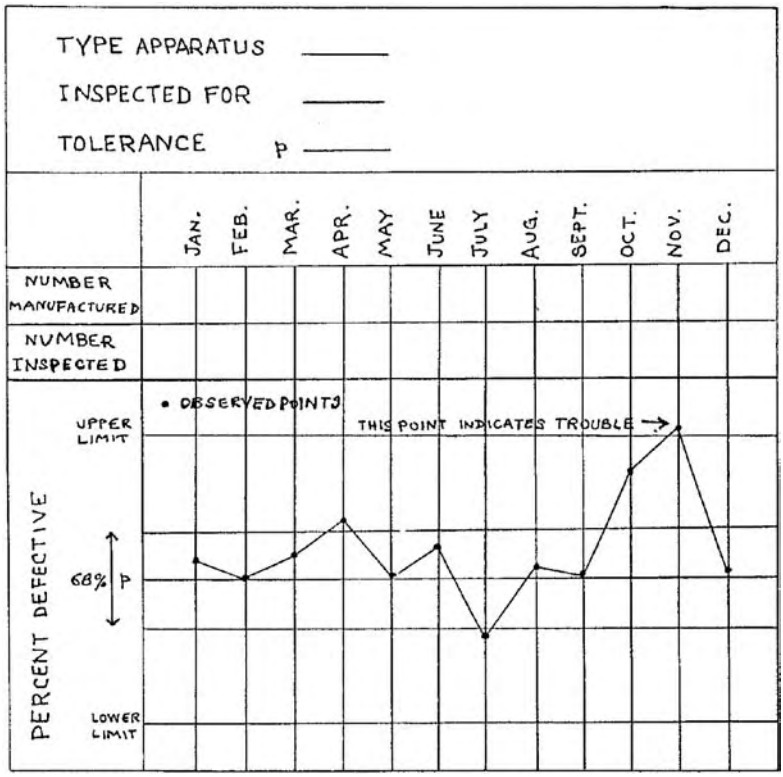


Figure 1. The first form of the control chart [Shewhart, 1924]. Property of Bell Telephone Laboratories, Murray Hill, New Jersey, USA.

The economic dimension mentioned in the title of the work arises for reasons of cost effectiveness. The identification and elimination of assignable causes of variation carry a cost, which must be comparable with the advantages resulting from an improvement in quality. In addition, physical phenomena introduce a natural variability (for Shewhart, all these physical magnitudes are of a statistical nature) which it is 'not reasonable' to seek and reduce. The quest for precision thus comes upon an obstacle both physical and economic, which imposes limitations on what the industry is capable of producing in terms of quality.

Beside the theoretical aspects Shewhart also conceived the cognitive tools which are an indisputable aid to the implementation of a course of action: these are 'control charts' (Figure 1). The control chart is a remarkable invention that depicts visually both the evolution of the characteristics of quality of production and the limits that it must not exceed. These limits represent the natural variability of the system. If they are exceeded at a given moment, this means that the system is no longer in the same stable state, that an external cause of variability has intervened to significant effect, and that an enquiry must be launched to identify it.

3 THE DIALOGUE BETWEEN MATHEMATICS AND INDUSTRIAL PRACTICE

The reader of Shewhart's book will perhaps be surprised by the mathematics to be found there, as it is closer to physics or engineering. Why then is it included in these 'landmarks of mathematics'?

One reason is that it is just this book that has opened the way for thousands of papers and reports on statistical process control. Another reason is that Shewhart puts mathematics to practical use without recourse to a brutal 'application' of formalisms or calculations. It institutes a dialogue between mathematical theory and experience by scrupulously emphasizing their respective limits to obtain the compromise necessary for action. This compromise is worked out from an 'economic' point of view, taking into account the cost of acquiring information and taking action in relation to the attendant benefits. Although these aspects are not developed very far, one can see in Shewhart the premises of the mathematics of decision-making.

On the mathematical plane, Shewhart invokes above all the following elements: a) the mathematical theory of statistical distributions (frequent use of the first four moments and various estimators); b) the identification of distributions from empirical observations, notably by the methods of Karl Pearson and the Gram-Charlier; c) the theory of sampling and distributions of sample statistics.

The works of Shewhart have a social dimension in their domain of action, industry: they introduce mathematical statistics into the factory, entrusting it to the working man with little education in mathematics. This is a *tour de force* accomplished by the control chart, which transforms into a visual metaphor the reasoning by statistical induction from samples.

It must be acknowledged that this work was accomplished on behalf of the large American telephone company AT&T, at the request and with the support of its directors, with the very rich human and material resources at the disposal of the departments of research and engineering. The legend 'Member of the Technical Staff, Bell Telephone Laboratories, Inc.' features prominently beneath the name of the author on the title page of the book. This involvement in business has given him access to real facts and industrial experience in creating a dialogue between statistical theory and practical experience.

Shewhart also had many exchanges with his colleagues, especially Harold F. Dodge, who also worked in statistical control but from another point of view (sampling inspection); and also those connoisseurs of probability and statistics, T.C. Fry and E.C. Molina. The works of Shewhart first appeared as the subject of 'Out-of-hours courses' at Bell Telephone Laboratories, whose technical services provided a valuable resource in making possible the realization of numerous graphs, tables and diagrams. Finally, the Bell Laboratories produced a scientific journal, *The Bell System technical journal*, which published several articles on these new methods.

Following in the footsteps of 'Student' (W.S. Gosset), Shewhart has thus contributed to the opening up of the field of industrial applications to mathematics. The new constraints suggested or imposed by the industrial terrain have stimulated the creativity of statisticians, leading to important theoretical advances. We recall, for example, that the statistical decision theory of Abraham Wald evolved from sequential analysis, which was itself a response

to a problem on the statistical control of quality raised during the Second World War by a Supply Corps officer [Wald, 1947].

4 PRESENTATION OF THE BOOK

The contents of Shewhart's book is summarized in Table 1. It consists of 25 chapters divided into seven parts, and three appendices. We shall take the parts in order, insofar as they correspond to a methodical progression through the various aspects of the material.

The introduction (Part I, three chapters) is very substantial. It comprises a synthesis of the whole, aimed at a scientific or cultivated audience. Shewhart gives a very systematic account of the hypotheses on which his method is based. These mainly concern the concept of 'constant system of causes', which he constructs by a method of abstraction based on equivalence relations between systems. He chooses examples of random systems known in the physical sciences for transferring methods of reasoning to industrial systems. He also describes the principle of the control chart and the various advantages that industry can reap from the statistical control of quality.

Part II ('Ways of expressing quality of product', six chapters) is an exposition of descriptive statistics adapted to the needs of industrial production. What are the different ways of presenting the facts expressing the quality of manufactured goods? What criteria should be used as evidence in making judgments? The order in which measurements are taken is essential when it is necessary to decide whether or not the system from which the series results is random. One chapter is devoted to tables, graphs and diagrams, another to analytic considerations, and a third to the expression of relations (correlations). This part recounts the work of a normalization committee created in 1929 to which Shewhart had contributed. The proposals of this committee were later adopted as standards by the American Standards Association [Littauer, 1950]. This is an example of Shewhart's contribution at the level of institutions of normalization for formalizing statistical facts.

Part III ('Basis for specification of quality control', three chapters) plumbs the theoretical depths of the new concept of control and proposes a characterization of 'maximum control'. This state is defined as 'the condition reached when the chance fluctuations in a phenomenon are produced by a constant system of a large number of causes in which no cause produces a predominating effect' (p. 151). A cause which dominates the others is called an 'assignable cause of Type I'.

Establishing that a system is in a state of control raises fundamental difficulties: such a state cannot be characterized in a positive way, as this would require observing its functioning over an infinite period of time. One can only make a hypothesis and submit it to experience. While lacking a sufficient condition for a state of control, one has the necessary condition that 'differences in the quality of a number of pieces of a product *appear* to be consistent with the assumption that they arose from a constant system of chance causes' (p. 146).

Part IV ('Sampling fluctuations in quality', four chapters) is devoted to sampling, both theoretical and practical. Shewhart utilises British mathematical statistics to establish relations between the original distribution and the sampling distributions for several estimators of central tendency and dispersion. For example, he draws up a table of

Table 1. Contents by chapters of Shewhart's book. The titles are quoted.

Chap.	**Page**	**Topics**
	vii	Preface, Table of contents.
Part I	1	*Introduction*.
1	3	Characteristics of a controlled quality.
2	8	Scientific basis for control.
3	26	Advantages secured through control.
Part II	35	*Ways of expressing quality of product*.
4	37	Definition of quality.
5	55	The problem of presentation of data.
6	63	Presentation of data by tables and graphs.
7	71	Presentation of data by means of simple functions or statistics.
8	85	Basis for determining how to present data.
9	99	Presentation of data to indicate relationship.
Part III	119	*Basis for specification of quality control*.
10	121	Laws basic to control.
11	145	Statistical control.
12	150	Maximum control.
Part IV	161	*Sampling fluctuations in quality*.
13	163	Sampling fluctuations.
14	174	Sampling fluctuations in simple statistics under statistical control.
15	214	Sampling fluctuations in simple statistics, correlation coefficient.
16	230	Sampling fluctuations in simple statistics, general remarks.
Part V	247	*Statistical basis for specification of standard quality*.
17	249	Design limits on variability.
18	262	Specification of standard quality.
Part VI	273	*Allowable variability in quality*.
19	275	Detection of lack of control in respect to standard quality.
20–21	301	Detection of lack of control.
Part VII	349	*Quality control in practice*.
22	351	Summary of fundamental principles.
23	376	Sampling, measurement.
24	404	Sampling.
25	418	The control program.
	425	Appendix I. Resultant effects of constant cause systems.
	437	Appendix II. Experimental results.
	473	Appendix III. Bibliography.
	493	Indexes of names and subjects. [End 501.]

coefficients for passing by simple multiplication from the observed mean of the standard deviation of the samples to the standard deviation of the distribution of the variable.

Shewhart also uses an experimental approach involving urns of controlled composition. First there is an urn whose composition simulates the normal law (discretized and represented by 998 chips marked according to the values of the random variable). This enables him to produce experimental results on samples and to verify the agreement between theory and experience. 4000 draws are made from this urn and reproduced in an appendix.

A second urn is made up of a population with rectangular distribution, that is, uniform between 0 and 1. A third represents a triangular distribution with maximum on the right. There also 4000 draws are made from each urn and the results published. They serve mainly to test the robustness of analytic formulae linking the original distribution and the sampling distributions where the former is not normal. These experiments lead to a conclusion which is very important in practice: samples of size $n = 4$ appear to be robust. Since that time the number 4 has featured in all the textbooks like a fetish (or sometimes 5, which makes the calculations easier). The underlying concept is that of 'rational subgroup' (see the next section).

In part V, Shewhart considers the formulation of production specifications ('Statistical basis for specification of standard quality', two chapters). In the first he studies from a statistical point of view the addition of tolerances, which is important in the design of products. He then proposes defining specifications of quality by means of a distribution, its mean and deviation. This represents another contribution to the scheme of drawing up new standards for industry.

Part VI ('Allowable variation in quality', three chapters) explains the theory of control charts and the criteria used in their construction (see the next section). Part VII ('Quality control in practice', four chapters) is a mixed bag. There is a chapter on the practical aspects of taking samples and another on measurements in physics aimed mainly at metrologists. The other two chapters reach very general and rather abstract conclusions, whereby Shewhart expresses his philosophy of control and of causality.

In the appendices, the author studies various constant systems of causes that are not normal, and their convergence towards normality as the number of causes increases. The results of the draws from the three types of box enable the reader to do his own experiments and to develop his intuition on the variability of samples. Finally, the well-annotated bibliography forms a valuable document for historians.

5 ALLOWABLE VARIABILITY IN QUALITY

The touchstone of Shewhart's method, being aimed at an operational result, is as follows: can one effectively reduce the variability of systems, and if so how? He distinguishes two types of situation.

The first type is that where the 'standard' of quality is given in advance, in the specification, in statistical form (mean and deviation, and form of the distribution). From the specified parameters of the distribution, one can calculate the limits between which the mean and deviation of a sample of size n drawn from this population must vary (with a

probability chosen close to 1). A sample for which the calculated statistic does not belong to the prescribed interval indicates with high probability that the system does not follow the specified distribution. One then pursues an enquiry in the physical world to find the source of this variation and eliminate it. The whole is subject to 'economic' considerations, by reason of cost-effectiveness.

The discussion on the choice of limits of control forms a particularly good illustration of the way in which Shewhart creates a dialogue between theory and experiment. The specified distribution serves essentially to demonstrate the existence of limits of optimal control from an economic point of view. But for the calculation of these limits, it is the experiment that predominates, with the intent of simplifying and uniformizing procedures. Shewhart finally recommends fixing the limits at 3σ (σ as standard deviation) on either side of the mean of the distribution. To justify this, he appeals to the very rapid convergence towards the normal law of a majority of the sample distributions. The choice of 3σ corresponds to a probability of 0.003 of making a bad decision, which is entirely acceptable in practice. He also makes appeal, when the law is arbitrary, to Chebychev's theorem, but without dwelling on the fact that the risk of a bad decision is thereby increased.

The second type of situation is much more interesting and innovative: bringing the system into a state of maximum control, of a kind where the norm is a property of the system itself and not just grafted on. The main difficulty in this case stems from the fact that one knows nothing definite about a possible statistical law of the phenomenon. It is necessary to work progressively.

To begin with, one must form 'rational subgroups' from the facts. This operation makes use of the knowledge of engineering of the system that one has at the outset. These subgroups serve to facilitate the discovery of the causes of variation (for example, by distinguishing the products according to the supplier). This concept, which is at an embryonic stage in the book, is made precise in [A.S.T.M., 1935, art. 4]: '*within* [a subgroup] the variations may be considered on engineering grounds to be due to nonassignable chance causes only, but *between* [subgroups] the differences may be due to assignable causes whose presence is suspected or considered possible'.

The next step is to construct a control chart for a suitable statistic (such as the mean, deviation or frequency of defectives). One estimates from samples the parameters of the distribution of this statistic, which serve to place the limits of control at $\pm 3\sigma$.

This first control chart enables one to detect assignable causes of variation. Having eliminated these, one repeats the process: rational subgroups, samples, calculation of new limits of control. These being tighter than the previous ones, reveal new assignable causes of variation. Continue in this way until a state of maximum control is reached or one judges that it is no longer economically viable to make further improvements. We mention, without going into details of the rather complex methods, the possibility of taking into account 'assignable causes of Type II' present in a system that seems to be in a state of control. Examples are causes of variation that have an effect en masse (correlated causes) and cannot be spotted individually.

6 RECEPTION OF THE BOOK

From the outset, the publication of Shewhart's book has concerned only statisticians and some specialists. Karl and Egon Pearson invited him to London in 1932 under the auspices of the British Standards Institution (BSI). The area was judged important enough to merit the creation of a special section of the Royal Statistical Society devoted to agricultural and industrial applications and a specialized series of their journal. In 1935, Egon Pearson published under the imprimatur of the BSI a textbook on Shewhart's work [Pearson, 1935] with certain improvements and modifications [Eisenhart, 1990].

Shewhart published only two books in his lifetime. The second is a collection of lectures delivered at an agricultural school in 1938, collected and edited by Deming, who had organized the lectures [Shewhart, 1939]. Here Shewhart developed his ideas on metrology, in a way that is disconcerting but undeniably profound. He was strongly influenced by the pragmatist philosophy of the logician Clarence I. Lewis and the operationalism of the physicist P. Bridgman.

It was, however, the economic regime of the War that ensured the wide dissemination of the statistical control of quality in industry, both in the United States and in Great Britain, under the impetus of governments who saw in it the technical and economic advantages of rationalizing production. Important educational programmes were set up.

After the War, the Marshall plan disseminated these methods of industrial organization in Continental Europe. But it was in Japan that they evoked the greatest response. Introduced in the postwar period by the MacArthur administration, they were gradually adopted by industrialists, who made them into a hobby-horse with world-famous success.

BIBLIOGRAPHY

A.S.T.M. 1935. 'Supplement B: "control chart" method of analysis and presentation of data', in *Manual on presentation of data*, Philadelphia: American Society for Testing Materials. [Repr. 1937.]

Bayart, D. 2000. 'How to make chance manageable: statistical thinking and cognitive devices in manufacturing control', in M.R. Levin (ed.), *Cultures of control*, Amsterdam: Harwood Academic Publishers, 153–176.

Deming, W.E. 1978. 'Shewhart, Walter A.', in W. Kruskal (ed.), *International encyclopedia of statistics*, New York: Wiley, vol. 2, 942–944.

Eisenhart, C. 1990. 'Shewhart, Walter Andrew', in C.C. Gillispie (ed.), *Dictionary of scientific biography*, New York: Scribner, vol. 18, Supp. II, 816–819.

Fagen, M.D. (ed.) 1975. *A history of engineering and science in the Bell system, the early years* (*1875–1925*), New York: Bell Telephone Laboratories.

Grant E.L. and Leavenworth R.S. 1972. *Statistical quality control*, New York: McGraw Hill (International Student Edition).

Hald, A. 1981. *Statistical theory of sampling by attributes*, New York: Academic Press. [*Industrial quality control*, *24* (1967), Shewhart special issue.]

Juran, J.M. 1997. 'Early SQC: a historical supplement', *Quality progress*, *30*, no. 9, 73–81.

Littauer, S.B. 1950. 'The development of statistical quality control in the United States', *The American statistician*, December, 14–20.

Pearson, E.S. 1935. *The application of statistical methods to industrial standardisation and quality control*, London: British Standards Institution.

Shewhart, W.A. 1924. 'Some applications of statistical methods to the analysis of physical and engineering data', *Bell system technical journal*, *3*, 43–87.

Shewhart, W.A. 1939. *Statistical method from the viewpoint of quality control*, Washington: The Graduate School, U.S. Department of Agriculture. Repr. New York: Dover, 1986.

Wald, A. 1947. *Sequential analysis*, New York: Wiley.

CHAPTER 73

VITO VOLTERRA, BOOK ON MATHEMATICAL BIOLOGY (1931)

G. Israel

In this pioneering book on mathematical Volterra studied aspects of the dynamics of animal populations. Some competition over results ensued with the statistician Alfred J. Lotka.

First publication. *Leçons sur la théorie mathématique de la lutte pour la vie*, Paris: Gauthier-Villars, 1931. vi + 214 pages.

Manuscript. Manuscripts in box 29 of the Volterra Archive at the *Accademia Nazionale dei Lincei*, Rome, Italy.

Photoreprint. Paris: Editions Jacques Gabay, 1990.

Related article: Wentworth Thompson (§64).

1 THE ORIGINS OF VOLTERRA'S INTEREST IN BIOMATHEMATICS

Vito Volterra (1860–1940) may be considered one of the greatest Italian mathematicians living between the 19th and the 20th centuries who also enjoyed great international prestige. He distinguished himself for his genius from a very early age: when he was only 13 he solved a restricted version of the three-body problem (compare §48). He was professor of mechanics and mathematical physics at Pisa, Turin and Rome. Nominated Senator of the Kingdom in 1905, he held numerous offices including that of President of the *Accademia dei Lincei* and founded new scientific institutions, including the Italian National Research Council (*Consiglio Nazionale delle Ricerche*). He was one of 12 Italian university professors who refused in 1931 to take an oath of allegiance to the fascist regime; this resulted in his being marginalized from national life and culminated in his total exclusion after the promulgation of the anti-Jewish race laws. His huge scientific production embraced central issues of mathematical physics, especially the theory of elasticity; of analysis, a field in which he is considered as one of the inventors of the theory of integral and integro-differential equations and of functional analysis; and applications of mathematics to biology.

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

In the early 20th century, Volterra may be considered as one of the few distinguished mathematical physical scientists to give credence to the use of mathematics in the non-physical sciences. In a famous lecture delivered in 1900 on the occasion of the opening of the academic year of Rome University [Volterra, 1901], he made an assessment of the applications of mathematics to the biological and social sciences. In his opinion, these applications ought to follow a mechanistic reductionist approach, consisting in 'transporting' into the new fields the methods that had been so successful in the mechanical and physical sciences. It is on this basis that he made his assessment of the results hitherto achieved. According to Volterra, the applications to economics had attained excellent results thanks to Léon Walras and Vilfredo Pareto, who had followed an explicitly mechanistic approach based on the methods of mathematical analysis (compare §41 on Jevons). Conversely, biology, with the exception of a few attempts such as the geometric model of the astronomer G.V. Schiaparelli, seemed to be dominated by statistical and probabilistic methods that Volterra considered to be a less important and non-rigorous branch of mathematics.

Despite the emphatic statement of the importance of pursuing such a programme of mechanistic mathematization, Volterra did not himself engage directly in this type of research in the following two decades. He concerned himself, in general terms, with mathematical economics, a field from which he later withdrew, perhaps because of the difficulties that arose out of his exchange of letters with Pareto. In the biological field, Volterra made no direct contributions until the mid 1920s, and it appears that he was not aware of the research of a mathematical nature carried out in that period, in particular by Alfred J. Lotka (1860–1949), an American statistician of Austrian origin with many interests ranging from population dynamics to chemical dynamics; and Ronald Ross (1857–1932), a British physician with a colonial background, a great expert in malaria, whose formulation of the dynamics of this disease in mathematical terms earned him the Nobel prize for medicine. On the other hand, there is no doubt that Volterra developed a strong empirical interest in the topic of the dynamics of animal populations and above all in fishery. This emerges from his activities on the Italian Oceanographic Committee, of which he was a founder member, promoting its coordination with similar initiatives at the European level. This interest was part of Volterra's general project to promote the applications of mathematics and was stimulated also by his scientific relations with Umberto D'Ancona (1896–1964), a distinguished zoologist in Italian scientific circles, and an expert in marine biology, who married his daughter Luisa.

2 THE PREMISES AND GENESIS OF THE BOOK: RESEARCH IN THE 1920S

The book with which we are dealing must be considered as the most important contribution to the mathematization of biology made during the first half of the century, together with [Lotka, 1925]. The origins of Volterra's direct intervention in the field of biomathematical research emerge clearly from his correspondence, manuscripts and direct testimony.

In late 1925 D'Ancona showed him the results of a statistical survey of fish populations in the Upper Adriatic sea that pointed to a curious phenomenon. As a general rule, the percentage of predatory fish in the total fish catch in several ports in the Upper Adriatic remained constant, while it displayed an appreciable increase during the period 1915–1918,

that is, the years during which Italy was at war and the naval conflict in the Adriatic had led to an interruption of fishing activities. D'Ancona suggested that the lull in fishing activities was the cause of the increase in the number of predators, and he asked Volterra to provide a mathematical proof of this. Volterra threw himself into the question and came up with a description of the interaction between prey and predators based on a simple mathematical model, which has since become famous and is known as the 'Lotka–Volterra equations'. Using this model he was able to verify D'Ancona's thesis, but he also drew from it the pretext for working out a much more extensive range of models to describe the interaction between any number of animal species in competition among themselves. He considered both the case of species in exponential growth and in limited (logistic) growth, and perfected his treatment by introducing delayed effects. To this he used the theory of memory systems (or 'hereditary systems') that he had developed in the field of the theory of elasticity and represented one of his most important results as a physico-mathematician. The differential equations involved were then replaced by integro-differential equations, a mathematical theory created by Ludwig Boltzmann, Emile Picard and by Volterra himself. These results were gathered together in an article published shortly afterwards [Volterra, 1926a] and a short summary of the more elementary results was published in the journal *Nature* [Volterra, 1926b] with a presentation by the well-known biologist D'Arcy Wentworth Thompson (§64), who had received it from an old friend of Volterra, the physicist Joseph Larmor.

It was clear that Volterra's interest was not so much in describing a series of scattered models as in giving form to the programme set out in the 1900 lecture: to introduce, at least for one branch of biology, a mechanistic approach based on the methods of mathematical analysis. Volterra actually defined his overall results as a 'rational mechanics of biological associations' that, as in physics, would be based on an experimental or at least empirical verification.

Volterra's articles, which preceded other publications in the years that followed, aroused widespread interest in the scientific world. Not only among his fellow mathematicians but also among biologists, His results, disseminated above all through the *Nature* article, aroused curiosity and interest. These biologists included the German Friedrich Bodenheimer, the zoologist of Canadian origin William R. Thompson, and the American entomologist Royal N. Chapman: Volterra kept up an interesting correspondence with them. Chapman played an important role in suggesting applications topics to him and in putting him in touch with his collaborator John Stanley and with another American entomologist, Samuel A. Graham. This led to the creation of a network of relations that developed further after the publication of Volterra's book. For further details and biographical information on the correspondents, see [Israel and Millán Gasca, 2002].

It must nevertheless be pointed out that Volterra's entry into the field also gave rise to some friction, as he had neglected to make any reference to Ross's work on malaria or to Lotka's book [Lotka, 1925], which indeed he did not know. Lotka wrote to *Nature* claiming that he had preceded Volterra in introducing his prey-predator equations in his book. This led to a (polite) dispute over priority [Lotka and Volterra, 1927]. It was easy for Volterra to show that his work had a much more ambitious aim than Lotka's and enunciated a much more general system of equations. Their correspondence reveals the totally different

perspectives of the two scientists, who never succeeded in establishing a fruitful dialogue [Israel, 1991, 1993].

In the meantime, Volterra ensured that his results were circulated in the mathematical world. In 1928, he was invited by the mathematician Emile Borel to Paris, a city in which he already felt at home, to give a series of lectures at the new *Institut Henri Poincaré* on the mathematical theory of biological fluctuations. The lectures were held in the winter of 1928–1929, and the task of collecting the text for possible editing was given to a young mathematician and former student at the *Ecole Normale Supérieure*, Marcel Brelot (1903–1987). This led to the idea of turning the text of the lectures collected by Brelot into a book to be published in the series *Cahiers Scientifiques*, directed by the mathematician Gaston Julia and published in Paris by Gauthier–Villars. In February 1929, Volterra wrote to D'Ancona to announce his decision to publish a book and to ask his opinion on three possible titles: 'Principes mathématiques de la lutte pour la vie', 'Théorie mathématique de la lutte pour la vie', and 'Principes mathématiques de biologie'. D'Ancona chose the second ('Mathematical theory of the struggle for life'), which Volterra followed; he resolutely rejected the third on the grounds that Volterra's studies involved only one sector of biology, ecology, and indeed only part of the latter. However, this was an indication of Volterra's ambition to open the way towards a general mathematization of biology.

3 THE WRITING AND CONTENTS OF THE BOOK

In order to appreciate the events surrounding the publication of the book something must be said about the person responsible for the final draft, namely Brelot. As was mentioned, he was a former student of the *Ecole Normale Supérieure*, where he had as fellow students mathematicians such as André Weil, Jean Dieudonné, Claude Chevalley and Henri Cartan, the founding nucleus of the 'Bourbaki' group. Although never actually a member, in his later career as a mathematician Brelot displayed such a passionate support for axiomatics as to be considered more Bourbakist than the Bourbakists. In 1931 he was engaged in writing a doctoral thesis under the guidance of Picard, a great friend of Volterra. It was due to the intervention of the latter and of Vessiot that Volterra facilitated Brelot being awarded a Rockefeller scholarship, which he utilized at the University of Rome under Volterra's guidance, and at the University of Berlin under Erhard Schmidt. It was only natural that the Roman period should be utilized for editing Volterra's book, even though Brelot did not like the climate or life in the Italian capital and made frequent trips back to Paris or to his native residence in the country at Boisseaux from where he wrote letters to Volterra.

According to the intentions of Volterra, Brelot was supposed to make an accurate transcription of his lectures, re-elaborating the mathematical part in all its details and in particular the proofs, adding appendices wherever necessary to make all the technical aspects easier for the reader to grasp. Furthermore, Volterra had a much broader readership that just mathematicians in mind: he aimed in particular at biologists. From a very early stage he involved D'Ancona, who was given the first drafts of the various parts of the book as soon as they were completed, and was regularly asked for his opinion on the terminology used, on all aspects of the biological side of the book, for the bibliographic references, and also to write some historical-bibliographical notes. In this way, Volterra aimed at putting

together a book that would satisfy two requirements: rigour and completeness of the mathematical treatment, so that it would represent a point of reference and departure for the development of research; and the richness of the references to biological issues, so as to involve biologists and stimulate as wide an interest as possible.

The task of holding together these two requirements proved more difficult than expected, especially because of the heterogeneous nature of the two collaborators and, in particular, because Brelot's mentality was quite different from Volterra's. Brelot was actually thinking of a book that, on the basis of a few simple biological ideas would be focused on 'rational research, calculations and mathematical theories', as he wrote in a letter to Volterra which forms part of the extensive correspondence between September 1929 to October 1930 that accompanied the writing of the book [Israel and Millán Gasca, 2002, ch. 4].

From this point of view, it is particularly interesting to examine the difference of opinion that occurred between Brelot and Volterra regarding the role of 'biological postulates' in the development of the theory. In Volterra's view, the process of mathematization should consist in formulating in mathematical terms hypotheses deemed to be plausible and then abandoning oneself to the tool of mathematical analysis, taking the process as far as possible, and only in the very end, to compare the results obtained with the initial hypotheses in order to verify whether any unsatisfactory aspects were due to any unrealistic aspects of the hypotheses themselves. Brelot, on the other hand, was in favour of introducing 'biological postulates' during the process in order to facilitate the mathematical treatment or to make it possible to obtain complete coherent partial results. He used this approach in the treatment of the case of a three-species ecosystem in which one species feeds off the second and the second feeds off a plant species. Volterra reacted somewhat energetically, calling upon him to modify his treatment saying that 'it is possible to go some considerable distance using mathematics and to do without postulates', as well as insisting on the need to maintain a distinction between the mathematical part and the biological part. In a series of phrases, subsequently deleted from the final version of the letter (dated September 1929) requesting Brelot to reappraise his approach, Volterra displayed all his annoyance, going as far as to say that 'postulates are mostly proposed by Satan in order to make us lazy' and developed in detail, also from a mathematical point of view, the treatment that ought to be followed by Brelot. The latter displayed a degree of resistance to Volterra's request, claiming that, however annoying, biological postulates represented the only way of ensuring rigour was maintained, that is, that the hypotheses were precisely defined, and went as far as to say that 'he felt some aversion to the approximation procedures' used by Volterra. The dispute ended in a verbal encounter of which we have no record but, judging from the result, a compromise was probably reached. Brelot attempted to limit the use of postulates, although he held out for his point of view quite stubbornly. Oddly enough, in a subsequent letter, he pointed out that 'a cultivated reader may find it annoying to have to leave the safe terrain of mathematical reasoning' (and use 'postulates'), 'but he can only blame the evil nature of the problem'. In other words, the blame for the unsatisfactory aspects of the treatment was not so much inherent in the latter as in biological reality itself.

Brelot's relative insensitivity to the empirical issues underlying the theory also explain the difficulty that he experienced in accepting D'Ancona's contributions. He considered D'Ancona's intervention to be superfluous and even incomprehensible. It is no coincidence

Table 1. Summary by Chapters and Sections of Volterra's book.

Ch.-Sect.: page	**Description**
v–vi: 1	Foreword.
I	*Coexistence of two species*.
I-I: 9	Two species competing for the same food.
I-II: 14	Two species one of which devours the other.
I-III: 27	Two species in the case of different mutual actions.
II	*Preliminary study of the coexistence of any number of species*.
II-I: 36	Species competing for the same food.
II-II: 38	First elements in the study of several species preying on each other.
II-III: 42	Case of an even number of species preying on each other.
II-IV: 58	Case of an uneven number of species preying on each other.
68	Mathematical note.
III	*Study of the coexistence of n species with more general hypotheses. Conservative and dissipative systems*.
III-I: 77	The coefficient of growth of each species living alone is allowed to depend on the number of individuals comprising it.
III-II: 96	Much more general theory.
III-III: 104	Conservative and dissipative associations.
III-IV: 131	Introduction of the hypothesis of variation of the external conditions.
135	Mathematical note.
IV	*On hereditary actions compared in biology and mechanics*.
IV-I: 141	Notion of inheritance of its mathematical translation.
IV-II: 159	Study of the coexistence of a predatory species and a preyed species in the hypothesis of an invariable linear heredity.
IV-III: 169	Hereditary energy in biology (preceding case with minor fluctuations) and in mechanics with a single parameter.
188	Mathematical note.
197	Conclusion, historical note, bibliography.
211–214	Contents.

that he found repugnant the idea of including the historical-biological note written by the latter in the book. According to Volterra's intentions this note was to have been an essential part of the book. Brelot accepted D'Ancona's intervention, pointing out that 'there was no harm in having the *apparent* collaboration of a professor of natural sciences' but insisted that it be contained in a separate note and placed at the end.

The contents of Volterra's book are summarised in Table 1. It consisted of four chapters. The first was dedicated to the problem of the coexistence of two species, the second to the

study of the coexistence of any number of species, which was then generalized in chapter three, in particular making a distinction between conservative and dissipative systems; chapter four introduced the notion of hereditary action in biology and the related technique of integro-differential equations. The book opened with an introduction that fully reflected Volterra's ideas and was concluded by D'Ancona's historical-bibliographical note. The book was accompanied by mathematical notes written by Brelot concerning elements of linear algebra, quadratic forms, and Volterra's integral and integro-differential equations.

4 THE PLACE OF THE BOOK AMONG VOLTERRA'S WORKS

With a process of construction such as this, it is quite clear that the book would not fully satisfy either its author or its editors, whether mathematicians or biologists. The book had the undeniable merit of providing a comprehensive and complete account of Volterra's research, of including it in the framework of research carried out on the subject, and of representing a useful handbook for anyone wishing to establish an exhaustive basis for further research. Its defects stemmed from the compromises made between the diverging ideas of the collaborators: Brelot's abstract and mathematics-oriented approach left heavy traces in the final draft despite Volterra's attempts to contain it. The book was too heavily biased in favour of mathematicians, and the over-abundance of technical jargon was likely to scare off biologists. Furthermore, Brelot's mathematical background revealed substantial gaps in the field of the qualitative analysis of differential equations, and so the treatment appeared rather old-fashioned and not to have taken on board more recent developments.

Volterra's dissatisfaction with the book, and in particular with its scant attraction for biologists, was expressed immediately and became the subject of an exchange of letters with D'Ancona, in which the blame for the defects was laid squarely on Brelot. This led to the idea of writing a new book addressed mainly to naturalists and thus pruned of all unduly complicated mathematical technicalities. This book, the result of an intense collaborative effort between D'Ancona and Volterra, was published in 1935 [Volterra and D'Ancona, 1935]. In Volterra's mind, the 1931 book represented the *rational phase* of the study of biological associations (corresponding to the status of rational mechanics in mathematical physics), while the book written with D'Ancona represented the development of the *applied phase*. In a third stage [Volterra, 1937] he was then to go on and develop the *analytical phase*, namely the formulation in variational terms of the mathematical theory of biological associations, corresponding to analytical mechanics [Israel, 1991].

5 THE BOOK'S RECEPTION AND ITS INFLUENCE ON BIOMATHEMATICAL RESEARCH

Volterra's book created quite a stir and gained him further scientific relations, particularly in the field of biology. The organic illustration of his results in a book aroused curiosity and stimulated attempts to compare theory with empirical reality. Furthermore, it imposed a certain direction on his scientific publications on the subject. It may be claimed that, whereas during the first phase interest in Volterra's work developed in the Anglo-Saxon world, it gradually gave way to new relations in the Continental European sphere. Significant relations were formed between Volterra and the eminent American zoologist Raymond

Pearl and the British zoologist Charles S. Elton (one of the founders of modern ecology) in the wake of the book's publication, although all the efforts to have the book published in English made by Elton, Chapman and D'Arcy Thompson were unsuccessful. In actual fact, the centre of Volterra's relations gradually shifted from the Anglo-Saxon scientific world to that of the Francophone and Soviet area. Volterra's research had strong repercussions on Soviet mathematicians, such as A.N. Kolmogorov who developed a generalization of Volterra's equations for the case of competing species. The Russian biologist Giorgii F. Gause showed considerable interest in Volterra's theories, engaging in an intense activity to verify them empirically [Gause, 1935].

However, it was above all in Paris that Volterra found two perceptive interlocutors who became his principal collaborators for the rest of his life: the Russian emigré geochemist Vladimir A. Kostitzin, and the professor of pharmacy Jean Régnier. Kostitzin, who had been a member of the 'Moscow school', was also an expert in the theory of integral equations and made a significant contribution to the theory, also of a mathematical nature [Kostitzin, 1934, 1937]. Régnier made available his laboratory for the purpose of carrying out empirical research on the growth of bacteria populations, with Kostitzin working on the theoretical side. However, the reductionist approach of the trio came into growing conflict with the dominant modelling approach. The war then either separated them physically or witnessed their deaths, and thus brought to a close that 20-year period known as 'the Golden Age of theoretical ecology' [Scudo, 1984].

For contemporary mathematical biology Volterra's 1931 book (and his biomathematical work as a whole) represents one of the most frequently cited references, as most ecosystem models are actually only re-elaborations and improvements of systems of equations that he enunciated. However, as a source for use the book is comparatively superficial, for most of its mechanistic scientific programme has been forgotten or abandoned.

BIBLIOGRAPHY

D'Ancona, U. 1926. 'Dell'influenza della stasi peschereccia nel periodo 1914–18 sul patrimonio ittico dell'Alto Adriatico', *Memorie del Regio Comitato Talassografico Italiano*, no. 126.

D'Ancona, U. 1939. *Der Kampf ums Dasein. Eine biologish-mathematische Darstellung der Lebesgemeinschaften und biologische Gleichgewichte*, Berlin: Borntraeger. [English trans.: *The struggle for existence*, Leiden: Brill, 1954.]

Gause, G.F. 1935. *Vérifications expérimentales de la théorie mathématique de la lutte pour la vie*, Paris: Hermann.

Israel, G. 1982. 'Volterra Archive at the Accademia Nazionale dei Lincei', *Historia mathematica*, *9*, 229–238.

Israel, G. 1991. 'Volterra's analytical mechanics of biological associations', *Archives internationales d' histoire des sciences*, *41*, 57–104, 307–352.

Israel, G. 1993. 'The emergence of biomathematics and the case of population dynamics: a revival of mechanical reductionism and Darwinism', *Science in context*, *6*, 469–509.

Israel, G. and Millán Gasca, A. 2002. *The biology of numbers. The correspondence of Vito Volterra on mathematical biology*, Basel: Birkhäuser.

Kingsland, S.E. 1985. *Modeling nature: episodes in the history of population ecology*, Chicago: University of Chicago Press.

Kostitzin, V.A. 1934. *Symbiose, parasitisme et evolution* (*étude mathématique*), Paris: Hermann.

Kostitzin, V.A. 1937. *Biologie mathématique*, Paris: Colin.

Lotka, A.J. 1925. *Elements of physical biology*, Baltimore: Williams & Wilkins. [Repr. as *Elements of mathematical biology*, New York: Dover, 1956.]

Lotka, A.J. and Volterra, V. 1927. 'Fluctuations in the abundance of a species considered mathematically', *Nature*, *119*, 12–13. [Two letters.]

Scudo, F.M. 1984. 'The "Golden Age" of theoretical ecology: a conceptual appraisal', *Revue Européenne des sciences sociales*, *22*, 11–64.

Volterra, E. 1976. 'Volterra, Vito', in *Dictionary of scientific biography*, New York: Scribners, vol. 14, 85–88.

Volterra, V. *Works*. *Opere matematiche, memorie e note*, 5 vols., Rome: Accademia Nazionale dei Lincei, 1955–1962.

Volterra, V. 1901. 'Sui tentativi di applicazione delle matematiche alle scienze biologiche e sociali', Discorso inaugurale, *Annuario della Reale Università di Roma*, 3–28. [Repr. in *Giornale degli economisti, (2) 23*, 436–458; in *Archivio di Fisiologia*, *3* (1906), 175–191; and in Volterra *Works*, vol. 3, 14–29. French trans. in *La revue du mois*, (1906), 1–20.]

Volterra, V. 1926a. 'Variazioni e fluttuazioni del numero d'individui in specie animali conviventi', *Memorie della Reale Accademia dei Lincei, (6) 2*, 31–113. [Repr. in. *Works*, vol. 5, 1–111.]

Volterra, V. 1926b. 'Fluctuations in the abundance of a species considered mathematically', *Nature*, *118*, 558–560.

Volterra, V. 1937. 'Principes de biologie mathématique', *Acta biotheoretica* (Leiden), *3*, 6–39. [Repr. in *Works*, vol. 5, 414–447.]

Volterra, V. and D'Ancona, U. 1935. *Les associations biologiques au point de vue mathématique*, Paris: Hermann (*Actualités scientifiques et industrielles*, no. 243). [Italian trans. by G. Israel, Rome: Teknos, 1995.].

CHAPTER 74

S. BOCHNER, LECTURES ON FOURIER INTEGRALS (1932)

Roger Cooke

This treatise marked a stage in the unification of harmonic analysis in the context of abstract integration theory. It contained the first publication of Bochner's famous theorem characterizing the Fourier transforms of positive measures.

First publication. *Vorlesungen über Fouriersche Integrale*, Leipzig: Akademische Verlagsgesellschaft, 1932 (*Mathematik und ihre Anwendungen in Monographien und Lehrbüchern*, vol. 12). viii + 229 pages.

Photoreprint. New York: Chelsea Publishing Company, 1948.

English translation. *Lectures on Fourier integrals. With an author's supplement on monotonic functions, Stieltjes integrals, and harmonic analysis* (trans. M. Tenenbaum and H. Pollard), Princeton: Princeton University Press, 1959 (Annals of Mathematics Studies, no. 42).

Russian translation. Lektsii ob integralakh Fur'e (trans. V.M. Borok, ed. Ya.I. Zhitomirskii, foreword by G.E. Shilov), Moscow: State Publishing House for Physics and Mathematics Literature, 1962.

Related articles: Fourier (§26), Riemann on trigonometric series (§38), Lebesgue and Baire (§59).

1 THE DEVELOPMENT OF THE THEORY OF THE FOURIER INTEGRAL

The need for a monograph on the subject of the Fourier integral arose from the development of analysis during the late 19th and early 20th centuries. This development brought about radical changes in both the meaning of the terms *function* and *integral* and in the understanding of what was meant by a mathematical proof. The context of set theory, within which the new theories of functions of a real variable and integration were developed,

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

raised entirely new questions in connection with the Fourier integral, so that a complete re-examination of its development needed to be undertaken. In order to make clear the issues addressed in this monograph, a brief sketch of the development of the Fourier integral during the 19th century will be helpful.

The problems that originally gave rise to the Fourier integral arose in the period 1806–1811 in work of P.S. Laplace, Joseph Fourier and S.D. Poisson [Poisson, 1814], and shortly thereafter in the work of A.L. Cauchy [Cauchy, 1817a, 1817b]. All of these men discovered the fundamental *Fourier inversion formula* that would nowadays be written

$$f(x) = \frac{2}{\pi} \int_0^\infty \int_0^\infty f(y) \cos(zy) \cos(zx)\, dy\, dz. \tag{1}$$

([Grattan-Guinness, 1972], and §26). In all three cases the discovery was connected with the classical equations known as the wave equation and the heat (diffusion) equation. The former equation was bequeathed to the 19th century by the 18th, and the latter was worked out by Fourier himself. The inversion formula contains two *integrals* and an unspecified *function* $f(y)$. The meaning of these two concepts underwent a metamorphosis during the 19th century, and that development led to an entirely different way of looking at the Fourier integral (§38).

More complicated problems were soon to follow. In studying heat conduction in a sphere Fourier found that it was necessary to use periodic functions whose frequencies were not all multiples of the same unit, but were solutions of the transcendental equation

$$nX/\tan(nX) = 1 - hX, \tag{2}$$

where X was the radius of the sphere and h a constant. With these more general series, it is not clear whether any function at all is represented outside the fundamental region. If anyone worried about what the series represented outside the interval of interest, no one seems to have written about it. In the physical world there are boundaries. At these boundaries certain functions, such as density and temperature have breaks. For that reason, mathematical explanations required analytic explanations whose validity was restricted to only some of the mathematically allowable values. The fact that sines and cosines repeat their values outside the fundamental period came to be accepted as irrelevant. Representations over a finite interval by functions that continue periodically seemed highly successful, and a discussion of the validity of such a representation was given by J.P.G. Dirichlet in 1829. In order to keep a consistent notation, we shall make a small update in the representation formula for Fourier series and state it only for even functions of period 2π, so as to bring out its analogy with (1). In this notation the Fourier inversion formula for series can be written

$$f(x) = \frac{1}{\pi} \int_0^\pi f(y)\, dy + \frac{2}{\pi} \sum_{n=1}^\infty \int_0^\pi f(y) \cos(ny) \cos(nx)\, dy. \tag{3}$$

The two representations (1) and (3) show a strong analogy, which one might possibly explain by their both having been generated by similar equations of mathematical physics. To a mathematician, such an analogy begs for a context in which the two things are both

manifestations of the same kind of underlying structure. But while the validity of (3) could be established under reasonable hypotheses (for example, it is valid at each point at which the function $f(x)$ has a derivative), the validity of (1) presented more subtle problems. The main source of such problems lay in the fact that the integral extended over an infinite interval, so that no merely local condition could assure convergence. In modern terms, both smoothness and reasonable decay at infinity were needed, and the latter threatened to impose such severe restrictions that the usefulness of the method for physical applications would be eliminated. This problem is an urgent one, no matter how the integral is interpreted. Before discussing the wider definitions of integrals, which altered the context of the inversion problem, we need to discuss the approaches to overcoming this difficulty.

In a study of wave motion in a fluid, Poisson [1823a, 1823b] also arrived at the Fourier integral formula in a rather strange-looking form, which, with slightly updated notation, we may write as

$$f(x) = \frac{1}{\pi} \int_0^\infty \int_{-\infty}^\infty f(\alpha) \cos a(x-\alpha) e^{-ka}\, d\alpha\, da. \tag{4}$$

The introduction of the 'convergence factor' e^{-ka} overcame the problem that seemed to arise so often, when the function defined by the inner integral did not decay quickly enough to guarantee convergence. The formula (4) is correct, if interpreted as the limit when k tends to zero through positive values. A very similar technique was used by N.H. Abel to justify formulas such as

$$\ln(2) = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots \quad \text{and} \quad \frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots, \tag{5}$$

which can be obtained by expanding the integrands in the following integrals as geometric series:

$$\int_0^1 \frac{1}{1+x}\, dx \quad \text{and} \quad \int_0^1 \frac{1}{1+x^2}\, dx. \tag{6}$$

Abel showed that the sum of the series was, as it appeared to be, the limit of the integrals from 0 to r as r increased to 1. The similarity to Poisson's technique can be seen by the substitution $r = e^{-k}$. This technique, which makes it possible to sum a series or evaluate an integral that, strictly speaking, diverges, is now called 'Abel–Poisson summation'.

As already mentioned, the concept of the integral also underwent a development during the 19th century, and of course any such development, because it affects the class of functions that can be integrated, forced a reinterpretation of the Fourier integral and a reexamination of its validity. The whole subject of calculus is founded on the identity of two apparently different things: a) the operation inverse to differentiation; b) the 'summation' of a continuous family of infinitely small products $f(x)\,dx$, where dx represents an 'infinitesimal increment' in the variable x. The fundamental theorem of calculus is based on the observation that $f'(x)\,dx$ is simply dy, where $y = f(x)$. The dubiousness of such reasoning and the needs of complex analysis led Cauchy to develop the integral in a different way, and a similar recasting of the definition was given by Bernhard Riemann in 1854, in connection with the study of trigonometric series representations. By the time of

Riemann's work a definition of continuity that is essentially modern had become accepted, and Riemann pointed out that under his definition some discontinuous functions could be integrated. It sufficed that, when the interval of integration was partitioned into sufficiently fine subintervals, the total length of the subintervals containing points where the integrand is discontinuous become arbitrarily small. (And, as Riemann showed, an even weaker condition suffices.) By allowing discontinuous functions to be integrated, Riemann had made it possible to form the Fourier integral of such a function (§39). But what properties would that integral have? Before that question was fully answered, the Riemann integral itself was superseded by a number of more general integrals, each of which demanded that the Fourier integral be reinvestigated.

2 NEW KINDS OF INTEGRALS

Under the impact of the new rigor demanded of series and integral representations and the powerful effect of the set theory created in the 1870s and 1880s by Dedekind, Cantor, and others—motivated in part by a desire to continue Riemann's excursion into the theory of trigonometric series representations—analysts began looking at the concept of the integral more and more closely as the 19th century drew to a close. New integrals appeared, associated with the names of A. Harnack, E. Borel, H. Lebesgue, A. Denjoy, and O. Perron. By far the most influential of these was the Lebesgue integral, created in all important essentials between 1899 and 1902 (§59). Two of its most profound effects were to be felt in probability, investigated by Borel, and trigonometric series representations, which Lebesgue recognized early and made the subject of a monograph. The beauty and importance of the Lebesgue integral were recognized early, and by 1907 E.W. Hobson (1856–1933) was including a discussion of this integral (in the form given by W.H. Young) in his *Theory of functions of a real variable*, which also included a chapter on applications to Fourier series [Hobson, 1907]. Meanwhile, the Lebesgue integral began to generate its own new questions, as F. Riesz introduced the classes now known as L^p, the spaces of measurable functions f for which $|f|^p$ is Lebesgue integrable, $1 \leqslant p \leqslant \infty$ (the space L^∞ consists of functions that are bounded on a set whose complement has measure zero). How the Fourier series and integrals of functions in these spaces behave became a matter of great interest, and a number of questions were raised, some of which required half a century to answer.

The difference between finite and infinite intervals and the extra hypotheses needed to ensure the convergence of the Fourier integral over an infinite interval opened up a gap in the understanding of the two types of transforms. A study of the validity of the Fourier integral formula was carried out by Alfred Pringsheim (1850–1941) in the article [Pringsheim, 1910]. He classified the hypotheses needed for the validity of the formula into two types, which he described as conditions in the finite region ('im Endlichen') and conditions at infinity ('im Unendlichen'). These two types of conditions are nowadays called *local* and *global* conditions. He pointed out that the local conditions could be traced all the way back to Dirichlet's work of 1829, but that 'a rather obvious backwardness reveals itself' in regard to the global conditions. In fact, he said, they

> seem in general to be limited to a relatively narrow condition, one which is insufficient for even the simplest type of application, namely that of absolute

> integrability of $f(\lambda)$ over an infinite interval. There are, as far as I know, only a few exceptions, such as those of A. Harnack, who, in the appendix to his German edition of J.A. Serret's *Textbook of Differential and Integral Calculus*, has given the following condition: $f(\lambda)$ must tend continuously to zero as λ increases without bound and must possess an absolutely integrable derivative $f'(\lambda)$.

Nowhere in this article did Pringsheim say whether he meant integrability in the traditional sense of Cauchy and Riemann or in one of the newer senses. He did refer to Lebesgue's *Leçons* (§59), however.

Actually the difference between local and global conditions is not so marked as Pringsheim implied. The pointwise convergence and summability proofs for both Fourier series and Fourier integrals require that the function have a certain smoothness or continuity at the point in question (local) and that it be bounded and integrable over the entire space (global). The difference is that for a finite interval a bounded function is automatically integrable, while such is not the case over an infinite interval. Therein lies the problem that makes integrals appear at first sight to be harder than Fourier series. In fact, right at the heart of the summation process lies a function of x called the *Dirichlet kernel*. Depending on an integer parameter n for a finite interval F and a real parameter R for an infinite one I, it is absolutely integrable over every F but not over any I. For a periodic function over an interval of lenght 2π it is

$$\sin(n+1/2)x \Big/ \sin x/2, \text{ while for the Fourier integral it is } \sin(Rx)/x. \tag{7}$$

Pringsheim also touched on a second issue mentioned above, namely the problem of what a series of trigonometric functions represents outside a finite interval if the frequencies of the representing functions are not all multiples of a fixed frequency. In his attempt to fill the lacuna in the global conditions he considered functions $F(x)$ given by absolutely convergent trigonometric series of the form $\sum c_\nu \cos(q_\nu x + \gamma_\nu)$, where $q_\nu \to \infty$ and $F'(x)$ is absolutely integrable. He mentioned that Fourier had considered such series, but he did not emphasize that, in contrast to Fourier, he needed to consider their values over the entire line, not just a finite interval. Such functions, a decade later, were to become the object of study in their own right, since they subsume all the periodic functions, as well as certain special functions considered earlier by P. Bohl and E. Esclangon. These more general functions, called *almost-periodic* functions, introduce another complication into any attempt to provide a unified theory of both series and integrals. Pringsheim did not show how the coefficients c_ν could be obtained from the function $F(x)$. Rather, he assumed that the coefficients were given in advance and that they defined the function. He did, however, say that

> It is perhaps worth noting that the set of functions $F(x)$ introduced into the Fourier integral formula by the preceding result is quite extensive [... and contains trigonometric series] progressing by completely *arbitrary* (that is, not necessarily integer) multiples of x, and hence can be, for example, of the type first considered by Fourier in his theory of heat conduction, which are characterized by the fact that the q_ν are the roots of a transcendental equation [...].

Pringsheim's negative remarks on the assumption of absolute integrability (quoted above) point up the problem to be overcome: the Dirichlet kernel, which allows a partial sum or integral to be expressed as an integral of the function being represented, is absolutely integrable in the case of Fourier series, but not in the case of the Fourier integral. Probably this complication accounts for the greater prevalence of discussions of Fourier series in textbooks in comparison with Fourier integrals. The lack of good reference works was a motive for the writing of both Bochner's *Vorlesungen* and a monograph of Norbert Wiener (1894–1964) that appeared about the same time. In the decade before these books appeared, both of these young men had made important contributions to the new area of almost-periodic functions, which was eventually to fit neatly along with Fourier integrals into a unified theory of Fourier analysis on a locally compact Abelian group.

As analysis was reaching new levels of abstraction in the first decade of the 20th century with the growth of functional analysis, a few analysts continued to write in the style of Karl Weierstrass and Leopold Kronecker, proving the existence of solutions to a problem by explicit construction. One such mathematician was Harald Bohr (1887–1951), brother of the famous physicist Niels Bohr and creator of the theory of almost-periodic functions. Bohr's original interest had been in number-theoretic questions, especially the Riemann hypothesis, on which he had worked together with Edmund Landau. This hypothesis concerns the Riemann zeta function, which can be represented for complex numbers z lying to the right of the number 1 in the complex plane by the Dirichlet series

$$\zeta(z) = \sum_{n=1}^{\infty} \frac{1}{n^z} = \sum_{n=1}^{\infty} \exp(-z \log n) = \sum_{n=1}^{\infty} c_n e^{-iy \log n}, \tag{8}$$

where the coefficients c_n in the case of this particular Dirichlet series depend on the real part (x) of z. For a fixed real part, this series is of the type considered by Fourier and Pringsheim. In contrast to the earlier cases considered by Fourier and Pringsheim, the function was given in advance, and the question as to how the coefficients were obtained for such a representation arose in earnest. The answer to that question posed a puzzle. Bohr discovered that for a function $f(x) = \sum c_\lambda e^{i\lambda x}$ the coefficient c_λ was given by

$$c_\lambda = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} f(x) e^{-i\lambda x}\, dx. \tag{9}$$

On the one hand, since periodic functions are a special case of this kind of series, this formula showed how to get the coefficients of the Fourier series of a periodic function by integrating over the entire line. On the other hand, it did not bring about the desired unity between series and integrals, since the formula involved an integral mean rather than an integral.

These investigations led Bohr to study the kinds of functions that can be approximated uniformly for all real x by general trigonometric polynomials, that is, by finite sums of the form

$$\sum c_\lambda e^{i\lambda x}. \tag{10}$$

A decisive step in the study of Fourier series of periodic functions had been Weierstrass's proof in 1885 that every continuous periodic function could be uniformly approximated by

a finite periodic trigonometric polynomial. Bohr now reversed the question in the context of these non-periodic series, asking which functions could be uniformly approximated by these more general finite trigonometric sums. He found the answer in a property that he aptly named *almost periodicity*. A function $f(x)$ is almost periodic if it has the following property, first stated by Bohr in 1923: For every $\varepsilon > 0$ there exists a positive number l such that every interval of length l contains at least one ε-*translate*, that is, a number τ such that $|f(x+\tau) - f(x)| < \varepsilon$ for all x. In Bohr's terminology, the ε-translates are 'relatively dense'. Although this language is quite in line with the modes of expression introduced by Cauchy, Weierstrass, and other 19th-century mathematicians, it leaves much to be desired as a starting point for the theory. For example, it is far from obvious that the sum or product of two almost-periodic functions is almost periodic. The analog of Weierstrass's approximation theorem, that is, the proof that an almost-periodic function can be uniformly approximated by general trigonometric polynomials, is one of the most difficult in all of analysis. The only known proofs are due to some outstanding 20th-century mathematicians (Bohr himself, Bochner, Wiener, Hermann Weyl, F. Riesz, and A.N. Bogolyubov). Bohr's proof was a good example of mathematics practiced according to strict Weierstrassian principles, full of explicit approximations. The route he followed was long and arduous, well described by Bochner's phrase in Bohr's obituary, that Bohr 'succeeded in proving the approximation theorem in his own fussy way, such as he did'.

Such was the state of affairs in classical Fourier analysis at the time when Bochner came to write his monograph. Fourier series of periodic functions had been well studied and were well adapted to the Lebesgue integral, but Fourier integrals did not fit so neatly into this theory, since the primary tool, the Dirichlet kernel, is not Lebesgue integrable. In addition, the new theory of almost-periodic functions, which included all periodic functions, was anomalous, since the Fourier coefficients were not obtained from the functions they represented by integration. A unified approach was the goal. As groundwork for such an approach, Bochner thought that a systematic exposition of the Fourier integral in the context of the now-dominant Lebesgue integral ought to be undertaken.

A single reference on which one could rely for the basic information was noticeably lacking. The dearth of expository material linking the Lebesgue theory of integration and the Fourier integral was apparent to Bochner's contemporary Wiener, who went from Cambridge, Massachusetts (MIT) to Cambridge, UK in 1932. That experience motivated him to write his own exposition of the subject, emphasizing the areas in which he had been one of the major contributors (the Wiener Tauberian theorem). In the preface to his monograph [Wiener, 1933] he explained what he had in mind:

> My original idea was of a rather comprehensive treatise, proceeding from the elements of Lebesgue integration through the L_2 theory of Fourier series to the Plancherel theorem, the Fourier Integral, the periodogram, and lastly, to theorems of Tauberian type. My impulse to write a book of this type arose from a dissatisfaction with the preponderant role of convergence theory in existing textbooks on the subject, and from the need for a treatment more in line with the extensive periodical literature.
>
> As far as my desire to write a book sprang from the need for a textbook to use in my course at the Massachusetts Institute of Technology, it has largely

> been dissipated by the recent appearance of a book on the Theory of Functions by Professor Titchmarsh. Several chapters of his book are devoted to the treatment of Fourier series from the modern point of view. Unfortunately—from my standpoint—he does not allot a great deal of space to the Fourier Integral and related matters. Thus, while there is now no need for the comprehensive treatise which I at first contemplated, there is need for a discussion of the Fourier Integral from the modern point of view. When Professor Titchmarsh's book has been in use for some five years, and has become the basis for higher instruction in Fourier series, it will be possible to treat the Fourier Integral in a thoroughgoing and coordinate way, but for the present we shall have to content ourselves with more fragmentary treatments.

The fact that two young men of very similar mathematical background (both had been working in almost-periodic functions, for example) chose to write an expository monograph on the same subject at the same time is intriguing, and a comparison of the two monographs is enlightening. We shall return to this topic after discussing the contents of the *Vorlesungen*.

3 THE AUTHOR

Salomon Bochner (1899–1982) was born near Krakov, Austria-Hungary, now part of Poland, and studied at the University of Warsaw. After receiving his Ph.D. at the University of Berlin in 1921, Bochner collaborated with Hardy and Littlewood in the United Kingdom and with Harald Bohr in Denmark. From 1924 to 1933 he was at the University of Munich, where he wrote the *Vorlesungen*. When Hitler came to power, he immediately emigrated to Britain. He soon received an offer from Princeton University, which he promptly accepted and acquired American citizenship. He taught at Princeton until the mandatory age limits in effect at the time forced his retirement in 1969. In that year he moved to Rice University in Houston, Texas, where he spent the remaining 13 years of his life.

As mentioned, just before writing the *Vorlesungen*, Bochner had been immersed in the theory of almost-periodic functions. In contrast to Bohr, Bochner made free use of the results of functional analysis. In his hands, the theory became much more transparent than Bohr had made it. For example, he defined almost-periodic functions as those functions whose translates form a conditionally compact set in the space of bounded continuous functions on the real line. To be sure, he did not use the description *conditionally compact*, but rather stated a condition equivalent to it: every sequence of translates of the function contains a uniformly convergent subsequence. It is easy to establish via the Ascoli–Arzelà criterion for compactness in the space of continuous functions on a finite interval that Bochner's definition is equivalent to Bohr's. From Bochner's definition it is immediate that the sum and product of almost-periodic functions are almost periodic. However, the benefits of this alternative definition go far beyond the ease of proving such elementary results. The definition reveals a fundamental symmetry about this class of functions that is obscured by Bohr's more traditional definition. The translations form a group operating on the almost-periodic functions, and the orbits under this group reveal important properties about the functions themselves. Through this symmetry group one can see connections

between the integrable functions, to which the Fourier integral applies, and almost-periodic functions. For example, Wiener showed that the translates of an integrable function $f(x)$ are dense in L^1 if and only if the Fourier integral of f never assumes the value 0. Thus, translations seemed to have an important relation to both integrable functions and almost-periodic functions, but a different one in the two cases.

In 1935 Bochner was to find a way of giving a unified discussion of almost-periodic functions (of a class even more general than those considered by Bohr, known as the *Stepanov* almost-periodic functions) with the Fourier integrals of integrable functions. The most general perspective on the problem, however, would not be gained until 1940, in the work of André Weil. All that, obviously, was in the future at the time the *Vorlesungen* were written.

4 BOCHNER'S BOOK

To restate the fundamental theorem about the Fourier series and the Fourier integral in the exponential form (introduced by Cauchy) that has nowadays generally replaced the use of sines and cosines, by the early 1930s there existed a considerable amount of literature on two kinds of Fourier inversion formulas: (a) the formula for integrals, which can be written as

$$f(x) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(y) e^{2\pi i z(x-y)}\, dy\, dz; \tag{11}$$

and (b) the formula for almost-periodic functions, which includes the case of Fourier series of purely periodic functions and can be written as

$$f(x) = \sum_{\lambda} \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} f(y) e^{-2\pi i \lambda(x-y)}\, dy, \tag{12}$$

where the limit of the integral is nonzero only for a countable collection of real numbers λ. The integral with respect to y in (11) and the limit of that integral in (12) necessarily converge if $f(y)$ is respectively absolutely Lebesgue integrable or almost periodic. The integral with respect to z in (12) and the summation over λ in (12), however, require either some special hypotheses about $f(y)$ or the introduction of some convergence factor in order to have a meaning.

The general direction in which to seek a unification of the two theories was suggested by the theory of representations of Lie groups, which underwent a spectacular development in the 1920s. It turned out that the functions that occurred as the entries in the matrices of a complete set of irreducible representations of a compact Lie group formed an orthogonal basis that could be used to represent functions, just like Fourier series. At that point, as K.I. Gross says [1994, 406–407]:

> The genius for bringing together these two seemingly unrelated themes [group theory and Sturm–Liouville theory] belongs to Hermann Weyl, who should be regarded as the father of modern harmonic analysis. The date of birth is 1927, and the official birth certificate is the remarkable paper by Weyl and his

Table 1. Contents by chapter of Bochner's book. The second column gives the pages in the German original, and the third column those in the English translation.

Ch.	Page	Page	Title
1	1	1	Fundamental properties of trigonometric integrals.
2	19	23	Representation and summation formulas.
3	39	46	The Fourier integral theorem.
4	63	78	Stieltjes integrals.
5	82	104	Operating with functions of the class $\Phi_0[L^1]$.
6	110	138	Generalized trigonometric integrals.
7	145	182	Analytic and harmonic functions.
8	169	214	Square integrability.
9	183	231	Functions of several variables.
	208	264	Appendix.
	219	281	Commentary [End 227].
		292	Monotone functions [Bochner, 1933]. [End 331.]

> student F. Peter [Peter and Weyl, 1927], in which the structure theory for the representations of a finite group is carried over completely to the context of compact Lie groups.

When Bochner succeeded in unifying Fourier series and integrals on Euclidean spaces in 1935, he was able to show immediately that his results could be applied to Lie groups, which are locally Euclidean spaces. One puzzle remained, however. The fact that a mean was needed, rather than an integral, to get the Fourier coefficients of almost-periodic functions, was not to be explained until harmonic analysis moved on to encompass more general groups in the years immediately following 1932.

The contents of Bochner's monograph is summarised in Table 1. It contains both a summary of the state of the subject in its new form and indications of alterations soon to come in the work of Bochner himself and others. In the foreword to the book the author explained the need for such a monograph:

> Since there is a wealth of material and as yet no book on the subject, I was forced to undertake the choice of material according to my own point of view. I set as my goal to develop the theoretical foundations for operations with Fourier integrals and their computational use. For that reason, alongside the general Fourier integrals, which have begun to be studied only recently, older things of a completely different type are also considered, for example, the evaluation of certain multiple integrals [...] which are not discussed in modern books and hence are unfamiliar to most younger mathematicians.

Bochner also included an appendix with statements of the main theorems about Lebesgue integration in several real variables, saying that 'this material has not yet found

its way into the textbooks'. That statement was not quite accurate when applied to the English literature, although it may have been in relation to the German literature. The third edition of Hobson's textbook had appeared in 1927, containing a thorough discussion of the Lebesgue integral in two variables, including the formula for change of variable. But where Fourier integrals were concerned, it was true that the connection with the Lebesgue integral had not yet been given an elementary exposition, even in one variable. The famous book of H.S. Carslaw included some applications and discussed the Lebesgue integral only in an appendix [Carslaw, 1930]. An English monograph on the Fourier integral, [Titchmarsh, 1937], was to be described by its author as 'a sequel to my *Theory of Functions*'. The latter book, published in 1932, had contained a systematic exposition of the Lebesgue integral.

Thus, in 1932, there was a clear field for a monograph on Fourier integrals in the context of Lebesgue integration. Bochner's book, however, was more than a systematic exposition of known results. In every chapter of it one can find applications that are new and interesting: applications to difference-differential equations, to integral equations, and many other areas. One chapter (the fourth) contained a result that was entirely new and was to have a profound influence on the subsequent development of the subject.

Along with the exposition, Bochner included a set of notes on the history of the subject. In his summary he said:

> The oldest textbook on Fourier Integrals (and in a certain respect the only one up to now) is the book *Analytische Studien*, Second Part, by O. Schlömilch [1823–1901], Leipzig, 1848.
> A. Pringsheim has made a worthy contribution to the history of Fourier integrals, especially in regard to the question as to how well it is justified to name them after J.J. Fourier, in his articles 'Über das Fouriersche Integraltheorem,' *Jahresbericht der deutschen Mathematiker-Vereinigung*, **16** (1907), 2–16, and 'Über die Gültigkeitsgrenzen für die Fouriersche Integralformel,' *Math. Ann.*, **68** (1910), 307–408. The latter [...] contains the first precise criterion for the validity of the Fourier integral formula and the Fourier integral theorem, which were further improved in a paper in *Math. Ann.*, **71** (1912), 289–298. The results of Pringsheim and later generalizations by other authors are all reproduced in the textbook of L. Tonelli [1885–1946] *Serie trigonometriche*, Bologna, 1928. We also note the book of E.W. Hobson *Theory of Functions of a Real Variable*, 2nd ed., Vol. 2, 1926 [...].
> The first precise result on the validity of the Fourier integral formula, which involved 'summability' rather than actual convergence, was stated long before Pringsheim by A. Sommerfeld [1868–1951] in his dissertation 'Über die willkürlichen Funktionen in der mathematischen Physik,' Königsberg, 1891.
> A very extensive collection of particular Fourier integrals can be found in the book *Fourier Integrals for Practical Applications* by George A. Campbell [1870–1954] and Ronald M. Foster [1896–1998].

Bochner dealt with the difficulty occasioned by the fact that the Dirichlet kernel is not absolutely integrable over the entire line by constructing parallel proofs of the main theorems to cover the case of a Lebesgue integrable function and the case of a function monotonically decreasing to zero. This parallel treatment of the two cases is pursued

throughout the development of the basic lemmas on Fourier integrals in Chapter 2 and is especially important in the main theorem of the subject, with which Chapter 3 opens. Stating the Fourier inversion formula as

$$\frac{1}{2}\big[f(x+0)+f(x-0)\big]=\frac{1}{\pi}\int_0^\infty d\alpha\int_{-\infty}^\infty f(\xi)\cos\alpha(\xi-x)\,d\xi \tag{13}$$

(he omitted the infinite limits of integration), he gave the following statement of the Fourier integral theorem:

> *A sufficient condition for the validity of* [(13)] *is that* $f(\xi)$ *be of bounded variation in a neighborhood of x and that one of the following conditions hold as* $\xi\to\infty$ *and* $\xi\to-\infty$:
> 1) $f(\xi)$ *is absolutely integrable,*
> 2) $\frac{f(\xi)}{\xi}$ *is absolutely integrable and tends monotonically to zero or, more generally, can be written in the form* $g(\xi)\sin(p\xi+q)$ *where* $g(\xi)$ *tends monotonically to zero.*

Bochner noted that it was permissible for one of these conditions to hold at $+\infty$ and the other at $-\infty$, and that the integral was to be understood as the Cauchy principal value in the second case. Chapter 3 then proceeds to develop the summability theory of Fourier integrals by various methods (Abel–Poisson, Gauss–Weierstrass, and others) and gives applications to the theory of Bessel functions and the evaluation of multiple integrals.

Chapter 4 ('Stieltjes integrals') provides the main claim of this work to 'Landmark' status. In the 20th century the groundbreaking new results in mathematics have nearly always appeared as research papers in journals, to be incorporated into expository monographs only later. Bochner, however, chose this monograph as the forum to reveal one of the most influential and profound results in Fourier analysis, a characterization of the Fourier–Stieltjes transforms of bounded nondecreasing functions. He defined a *distribution function* to be a nondecreasing bounded function whose value at each point is the average of its right- and left-hand limits. For such a function $V(\alpha)$, the *Fourier–Stieltjes transform* is defined as the integral

$$F(x)=\int e^{-i\alpha x}\,dV(\alpha). \tag{14}$$

Bochner denoted the set of all such transforms B. In Theorem 23 he gave the following characterization of these transforms: '*In order for a function to belong to the class B, it is necessary and sufficient that it be positive-definite*'.

Bochner defined a positive-definite function to be a continuous function $f(x)$ that is *Hermitian*, meaning that $f(-x)=\overline{f(x)}$, where the bar denotes complex conjugation, and also has the property that for any real numbers $x_1,\dots,x_m$ and any complex coefficients $\rho_1,\dots,\rho_m$, the following inequality holds:

$$\sum_{\mu=1}^m\sum_{\nu=1}^m f(x_\mu-x_\nu)\rho_\mu\overline{\rho}_\nu\geqslant 0. \tag{15}$$

(The double sum is necessarily a real number because of the Hermitian property.) Such functions had been considered by Gustav Herglotz two decades earlier in connection with Fourier series and the famous problem of characterizing the moments of a Stieltjes integrator. The definition just given had been formulated in 1923 by M. Mathias, who had investigated some of the properties of positive-definite functions on the real line.

Nevertheless, it was primarily Bochner who realized the importance of this concept. Although his name is justifiably attached to a number of results in various areas of mathematics, it is this theorem above all that analysts tend to mean when they say 'Bochner's theorem'. It was generalized by A. Weil and D.A. Raikov to the setting of locally compact groups and plays an important role in understanding the behavior of Fourier–Stieltjes transforms. The same theorem was discovered about this time by the Soviet mathematician A.Ya. Khinchin and published in the *Bulletin of Moscow State University* in 1937. For that reason it is referred to in Soviet literature as the Bochner–Khinchin theorem. Considering all its prefigurations and generalizations, Loomis [1953] called it 'the Herglotz–Bochner–Weil–Raikov theorem'. Bochner gave it an immediate application, providing a new proof of the Parseval relation for almost-periodic functions. Reversing the usual order of presentation for such new results, Bochner soon wrote a research paper containing this result and a number of others [Bochner, 1933].

In July 1932, when he wrote the words quoted above, Wiener probably did not know of Bochner's book, although the two men certainly knew each other. Each cited work of the other in his monograph. Bochner devoted an entire section (article 9 of Chapter 2) of the *Vorlesungen* to a generalization of a formula of Wiener (implicit in Wiener's work, as he said). While proving the Parseval relation for almost-periodic functions in his monograph, Wiener cited Bochner's work on almost-periodicity. As this last sentence indicates, applications to almost-periodic functions, an area in which both authors had worked, are a common topic in the two books. In neither book, however, are they a major part. They occupy more of Wiener's book than of Bochner's, but even there almost-periodic functions form only a part of one chapter. Of necessity, both books discuss the fundamental results of the theory, such as Plancherel's theorem. However, the proofs in the two cases are so different as to constitute almost a difference in kind. The two men seem to have looked at the subject from entirely different points of view. This difference did not preclude their using each other's results in their own work. For example, Bochner used Wiener's celebrated theorem that the reciprocal of a nonvanishing function with an absolutely convergent Fourier series also has an absolutely convergent Fourier series to prove a very delicate theorem on absolute convergence of multiple Fourier series in [Bochner, 1936].

5 THE AFTERMATH: ABSTRACT HARMONIC ANALYSIS

The year following the appearance of Bochner's monograph brought the last and greatest paper of Alfred Haar (1885–1933), which contained a proof of the striking fact that every locally compact topological group possesses a translation-invariant measure, known in his honor as *Haar measure* [Haar, 1933]. Haar measure made it possible to generalize the Fourier integral to the context of any such group. Haar's proof of the existence of a translation-invariant measure inspired John von Neumann, the following year, to prove the

existence of a translation-invariant *mean*, analogous to the Bohr mean. This mean allowed von Neumann to generalize the notion of almost-periodicity to functions on any group. He simply showed that the convex set spanned by the translates of such a function contained a unique constant function, equal to the mean. Von Neumann's construction looked like a radical step away from the concept of a topological group, since it seemed to ignore the topology completely. Whereas the ordinary dual group consisted of continuous homomorphisms from the group into the circle, von Neumann was considering arbitrary homomorphisms, whether continuous or not. That fact finally provided the clue that put almost-periodic functions in the proper perspective. Considering both continuous and discontinuous homomorphisms amounted to putting the discrete topology on the dual group, and hence making its dual group compact. Since the underlying group is imbedded in the dual of the dual of the group with the discrete topology, almost-periodic functions could be seen as the restrictions of functions that are continuous on the dual of the discrete dual. This compact group is now called the *Bohr compactification* of the group, and von Neumann's (and Bohr's) mean was simply the integral with respect to the Haar measure on the Bohr compactification. In this way complete unity was at last achieved, and all forms of (commutative) Fourier analysis were special cases of functions on a locally compact group and its dual.

Although harmonic analysis became *abstract* harmonic analysis in order to unify Fourier series and integrals, the classical cases of periodic or almost-periodic functions and Fourier integrals of integrable functions on Euclidean space remained a topic of special interest, both in its own right, because of the specialized theorems and questions it generated, and as background for the more abstract study. A generation after the publication of Bochner's book, mathematicians were still turning to it for information and inspiration. In 1959 two faculty members at Cornell University, Morris Tenenbaum and Harry Pollard (co-authors of a well-known textbook of differential equations), published with Princeton University Press an English translation of the *Vorlesungen* and included [Bochner, 1933]. As they said,

> our main purpose was to make generally available to the present generation of group-theorists and practitioners in distributions the historical and concrete problems which gave rise to these disciplines. Here can be found the theory of positive definite functions, of the generalized Fourier integral, and even forms of the important theorems concerning the reciprocal of Fourier transforms.

Three years later, in 1962, the English translation formed the basis of a Russian translation. In his foreword G.E. Shilov noted the same points as Tenenbaum and Pollard in justifying the publication of a 30-year-old work, pointing out that Bochner's work on generalized trigonometric integrals anticipated part of the theory of distributions of Laurent Schwartz, namely the part relating to the Fourier transform of slowly increasing functions. After listing the great changes in harmonic analysis over the preceding generation due to the application of the theory of analytic functions of a complex variable, the Gel'fand theory of representations of Banach algebras and the theory of generalized functions due to Gel'fand, Schwartz, and others, Shilov concluded: 'All that being said, in its wealth of specific material, its "classic" character (in the best sense of that word), and in its absolute accessibility, Bochner's book retains its full value even for the present time'.

BIBLIOGRAPHY

Annaratone, S. 1997. 'Les premières démonstrations de la formule intégrale de Fourier', *Revue d'histoire des mathématiques*, *3*, 99–136.

Bochner, S. 1933. 'Monotone Funktionen, Stieltjessche Integrale und harmonische Analyse', *Mathematische Annalen*, *108*, 378–410.

Bochner, S. 1936. 'Summation of multiple Fourier series by spherical means', *Trans. Am. Math. Soc.*, *40*, 175–207.

Carslaw, H.S. 1930. *Introduction to the theory of Fourier's series and integrals*, 3rd ed., London: Macmillan. [Repr. New York: Dover, n.d.]

Cauchy, A.-L. 1817a, 1817b. 'Sur une loi de réciprocité qui existe entre certaines fonctions' and 'Seconde note', *Bulletin des sciences, par la Société Philomatique de Paris*, 123–124, 178–181. [Repr. in *Oeuvres*, ser. 2, vol. 2, 223–232.]

Dieudonné, J. 1971. 'Histoire de l'analyse harmonique', unpublished manuscript.

Grattan-Guinness, I. with Ravetz, J.R. 1972. *Joseph Fourier, 1768–1830*, Cambridge, The MIT Press.

Gross, K.I. 1994. 'Harmonic analysis', in I. Grattan-Guinness (ed.), *Companion encyclopedia of the history and philosophy of the mathematical sciences*, London: Routledge, 395–418.

Haar, A. 1933. 'Das Massbegriff in der Theorie der kontinuerlichen Gruppen', *Annals of mathematics (2) 34*, 147–169. [Repr. in *Gesammelte Arbeiten*, 600–622.]

Hobson, E.W. 1907. *The theory of functions of a real variable*, 1st ed., Cambridge: Cambridge University Press. [2nd ed. 2 vols., 1921–1926.]

Loomis, L. 1953. *Abstract harmonic analysis*, Princeton: Van Nostrand.

Peter, F. and Weyl, H. 1927. 'Die Vollständigkeit der primitiven Darstellungen einer geschlossenen kontinuierlichen Gruppe', *Mathematische Annalen*, *97*, 737–755.

Poisson, S.-D. 1814. 'Mémoire sur les intégrales définies', *Bulletin des sciences, par la Société Philomatique de Paris*, 185–188.

Poisson, S.-D. 1823a, 1823b. 'Mémoire sur la distribution de la chaleur dans les corps solides' and 'Seconde mémoire', *Journal de l'Ecole Polytechnique*, cah. 19, *12*, 1–144, 290–403.

Pringsheim, A. 1907. 'Über das Fouriersche Integraltheorem', *Jahresbericht der Deutschen Mathematiker-Vereinigung*, *16*, 2–16.

Pringsheim, A. 1910. 'Über neue Gültigkeitsbedingungen für die Fouriersche Integralformel', *Mathematische Annalen*, *68*, 367–408.

Pringsheim, A. 1912. 'Nachtrag zu der Abhandlung "Über neue Gültigkeitsbedingungen für die Fouriersche Integralformel"', *Mathematische Annalen*, *71*, 289–298. [Contains the correction of an error in [1910], pointed out by W.H. Young, and some extensions of the result.]

Titchmarsh, E.C. 1937. *Introduction to the theory of Fourier integrals*, Cambridge: Cambridge University Press.

Tonelli, L. 1928. *Serie trigonometriche*, Bologna: Zanichelli.

Wiener, N. 1933. *The Fourier integral and certain of its applications*, Cambridge: Cambridge University Press.

CHAPTER 75

A.N. KOLMOGOROV, *GRUNDBEGRIFFE DER WAHRSCHEINLICHKEITSRECHNUNG* (1933)

Jan von Plato

In this short book Kolmogorov laid out the foundations of probability theory in terms of set and measure theory, bringing into definitive form a line of thought among some probabilists of the past three decades. His handling of conditional probabilities and of infinite fields of probability was especially significant.

First publication. Berlin: Springer, 1933 (*Ergebnisse der Mathematik*, vol. 2, no. 3). ix + 62 pages.

Reprint. Same publisher, 1973.

English translation. Foundations of the theory of probability (trans. N. Morrison), New York: Chelsea, 1950. [2nd ed. 1956.]

Related articles: Laplace on probability (§24), Lebesgue and Baire (§59), Gödel (§71), Hilbert and Bernays (§77).

1 INTRODUCTION

The *Grundbegriffe der Wahrscheinlichkeitsrechnung* ('Fundamental concepts of probability') by Andrei Kolmogorov (1903–1987) is the book that has become the symbol of modern measure-theoretic probability theory, its year of appearance 1933 being seen as a turning point that made earlier studies redundant. The idea of a measure-theoretic foundation of probability was almost as old as measure theory itself, and it had been repeatedly presented and used in the literature prior to 1933. Therefore, the mere idea was not the reason for the acceptance of Kolmogorov's measure-theoretic approach, but rather what he achieved by its use. The two essential mathematical novelties of the *Grundbegriffe* were the theory of conditional probabilities when the condition has probability zero, and the general theory of random or stochastic processes. For works not explicitly referred to below, see the bibliography of [von Plato, 1994].

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

Kolmogorov's mathematical education was in the spirit of set theory and modern real analysis of the school of N.N. Lusin (1883–1950). His first paper, written in 1922 but published in 1928, deals with what is today called descriptive set theory. In his first publications he studied the properties of Fourier series, giving an example of a series that diverges except for a set of measure zero [Kolmogorov, 1923]. In [1926] he gave an everywhere divergent series. In his [1925] he published the first of his early papers on logic and foundations of mathematics; a belated English translation appeared only in 1967. It is the first publication ever on a constructive system of logic, and Kolmogorov's main aim in it was to 'save' classical mathematics by showing that its inferences are formally acceptable. He gives a translation of classically provable formulas to constructive ones, an invention usually attributed to Kurt Gödel and Gerhard Gentzen, but besides the formal results it also contains a general intuitionistic approach to the philosophy of mathematics. Passages from the *Grundbegriffe* (hereafter, 'GW') remind us that Kolmogorov maintained his intuitionist philosophy presented in his [1925]. Referring to sets (events in probabilistic terminology) that are infinite unions of other sets, he writes that 'we consider these sets in general only as ideal events to which nothing corresponds in the world of experience. However, if a deduction uses the probabilities of such events and if it leads to the determination of the probability of a real event, this determination will obviously be unobjectionable also from an empirical point of view' [p. 16].

2 THE BACKGROUND OF THE *GRUNDBEGRIFFE*

In 1900 David Hilbert (1862–1943) presented his list of mathematical problems at the International Congress of Mathematicians in Paris (§57). Hilbert's sixth problem is, following the example of his *Grundlagen der Geometrie* one year earlier, to treat axiomatically those physical disciplines in which mathematics plays a predominant role (§55). These are in the first place the calculus of probability and mechanics. Hilbert added that it would be desirable to have, together with the logical investigation of the axioms of probability theory, a rigorous and satisfying development of the methods of determining averages in physics. This goes specifically for the kinetic theory of gases. Early attempts at the axiomatization of probability include such forgotten names as Laemmel, Broggi, and a few others.

Kolmogorov's suggested solution to Hilbert's sixth problem had also later predecessors; he himself mentions Richard von Mises and Sergei Bernstein as exponents of axiom systems with interests different from his. In these, the concept of probability is a defined notion, and the attempt is 'to establish a connection as close as possible to the empirical origin of the concept of probability' (p. 2). However, Kolmogorov concludes that for the sake of simplicity of the theory, 'it seems most appropriate to axiomatize the concepts of a chance event and its probability' (p. 2). Kolmogorov wanted to follow the example of Hilbert's *Grundlagen der Geometrie* in the questions of formalization. Probability theory is to be formalized in exactly the same abstract way as geometry or algebra. As a consequence, the formalism has several other interpretations in addition to the one from which it grew. Thus, probability theory can be applied to cases which 'do not have anything to do with the concrete sense of the notions of chance and probability' (p. 1). Behind this statement there is an application of probability to a purely infinitary situation. The idea of

basing probability theory on measure theory is by no means original in Kolmogorov, as we have mentioned. He says himself that after Lebesgue, 'the analogy between the measure of a set and the probability of an event, as well as the integral of a function and the mathematical expectation of a random quantity, lay at hand' (p. iii). Subsequently, Maurice Fréchet formulated measure theory in an abstract way, so as to make it independent of its origins as a generalization of geometric measure. Kolmogorov says that this abstraction made it possible to found probability on measure theory, and that 'the construction of probability theory according to these points of view has been current in the appropriate mathematical circles' (p. iii).

Two works precede the measure-theoretic axiomatization of the *Grundbegriffe* [Kolmogorov, 1929, 1931]. In the latter, there was a physical motivation for building a theory of probability, namely the need to handle schemes of statistical physics in which time and state space are continuous. Probability was introduced as a σ-additive measure over the state space, as the transition probability $P(t_1, t_2, x, A)$ for going from state x at time t_1 to the set of states A at time t_2. It was supposed to be a measurable function with respect to x, with expected values of random variables defined as Stieltjes integrals. The theory of continuous processes was built directly upon the model of classical physics. The state space of a classical system is usually a subset of a real space R^n. The present state of a system and its dynamical law determine its future behaviour. In a probabilistic generalization, the present state determines a probability distribution over future states, leading to the notion of a Markov process. A paper preceding the random processes of 1931 by only two years in publication and much less in writing, namely [Kolmogorov, 1929], contains nothing of the physically oriented motivations. Titled 'General theory of measure and the calculus of probability', it tries to show the possibility of 'a completely general and purely mathematical theory of probabilities'. Further, 'finding out from the formulation of probability theory those elements which condition its inner logical structure and do not relate at all to its concrete meaning, is sufficient for such a theory'. The theory is consequently wider in its range than a calculus of probability which is only meant to deal with chance phenomena, for the former extends to the realm of pure mathematics. Kolmogorov mentions as an example the distribution of the digits of a decimal expansion, a result found with the help of the formulas of the calculus of probability, but not involving any concrete notion of chance. On the relation between measure theory and probability he says that 'the general concept of measure of a set contains the concept of probability as a special case'. Therefore the results of probability theory concerning random variables are special cases of results on measurable functions.

The concept of *independence* is central in the application of probability theory to pure mathematics. Kolmogorov says this concept had never been formulated purely mathematically before. Obviously, one need for such a definition is the independence of the digits of a decimal expansion. The thought being that arithmetic sequences follow some law or other, their independence property has to be saved from the domain of chance by a purely formal mathematical definition. The conditions for a finitely additive probability are laid down as axioms, and denumerably additive measures are signalled out by the term 'normal'. In a note added to the reprinting, Kolmogorov [1986, 472] mentions that this early work did not yet contain the set-theoretic notion of conditional probability. One could speak of a set-theoretic foundation of the whole of probability theory only after conditional probabilities

as well as distributions in infinite product spaces had been incorporated, he says. These are the two mathematical novelties of the *Grundbegriffe*, the book which established that set-theoretic foundation.

3 AXIOMS FOR FINITARY PROBABILITY THEORY

The book proper starts with the axiom system for a finite set of events. In view of Kolmogorov's position on foundations of mathematics, this is very natural. The famous axioms go as follows (p. 2):

There is a set E of *elementary events* $x, y, z, \ldots$. There is a family of subsets $\mathscr{F}$ of E, the members of which are called 'chance events'.

I. $\mathscr{F}$ is a field of sets (that is, closed with respect to unions, intersections, and complements).

II. $\mathscr{F}$ contains the set E.

III. To each set A of $\mathscr{F}$, a non-negative real number $P(A)$ is attached. This number $P(A)$ is called the probability of the event A.

IV. $P(E) = 1$.

V. If A and B are disjoint, $P(A \cup B) = P(A) + P(B)$.

We shall first review the place of foundational questions in the book.

First of all, Kolmogorov does not offer a formalization of probability in the strict sense of the word, but an *informal axiomatization* within intuitive set theory. It is of course straightforward to give a strict formalization, by giving the axioms in a formalized system of set theory. Set theory and the measure-theoretic way of building up the theory of real functions were the kind of mathematics in which he was educated. The reference for set theory in the book is the short version of Felix Hausdorff's *Mengenlehre* [Hausdorff, 1927]; the presentation of measure-theoretic probability in the first edition of 1914 was left out in this shorter version. Kolmogorov shows first that his axiom system is *consistent*. In logical terms, he gives an *interpretation* for the formal axioms. An interpretation, or a *model*, consists of a *domain* D, the set of objects the interpretation talks about, and a set of *relations* F. These latter specify the functions and relations of the domain that correspond to the functions and relations of the formal axioms. This correspondence has to be such that the relations which interpret the formal notions, are fulfilled in the domain. Corresponding assertions about the relations are *true in the model*. Specifically, the axioms correspond to relations that hold in the model. A contradictory axiom system is one that has no models. Conversely, if an axiom system has at least one model, it is non-contradictory. The model that Kolmogorov puts up is very simple: for E, take any set $\{x\}$, so the domain interpreting $\mathscr{F}$ is the set $\{\emptyset, E\}$. Defining the function P by $P(E) = 1$ and $P(\emptyset) = 0$, it is easy to see that this interpretation fulfils the axioms.

Next Kolmogorov notes that the axiom system is, as he says, 'incomplete' ('unvollständig'). Some caution is in order here. At the time of the writing of Kolmogorov's book, work on foundations of mathematics was in full progress. The great name in foundational

studies was of course Hilbert. His 'metamathematics' had as its objects the formalization, proof of consistency and completeness, and creation of a decision method for mathematics (§77). In more modern terms, the completeness of an axiomatic system can be explained as follows. An assertion is *logically true* or valid if it holds in all possible models; there is no counter-example. An axiom system is *complete* if all true assertions can be formally proved in it. The completeness of predicate logic was proved by Kurt Gödel in 1930. Next, an axiom system is *incomplete* if there is a true assertion for which there is no formal proof within the system. Gödel's sensational incompleteness theorem of 1931 shows that there are true assertions of formalized arithmetic for which there is no such proof (§71).

The notion that Kolmogorov is after in the passage under discussion is another one: that of the *categoricity* of an axiomatization. An axiomatization is categorical if all its models are isomorphic. In his book Kolmogorov says that his axiomatization of probability is incomplete because in different problems of probability theory one considers different fields of probability (p. 3). Obviously, the intention is that there are non-isomorphic fields so that the axioms of probability do not characterize their possible interpretations in a categorical way. A further metamathematical question concerns the mutual *independence* of the axioms. If there is an interpretation which makes all but one of the axioms of some system true, there cannot be any logically correct deduction of that axiom from the others. Kolmogorov seems to take the independence of his axioms for obvious. Indeed, simple considerations show the independence: not all measures all normalized, so that there cannot be any deduction of axiom IV. As there are genuine subadditive measures, axiom V is independent, and so on. In the treatment of infinite fields of probability, an additional axiom VI is posed. Kolmogorov shows that it is independent of the other axioms. However, under the additional assumption of a finite field of probability, it becomes derivable (p. 14).

4 THE APPLICATION OF PROBABILITY

The sense of probability that Kolmogorov endorses is addressed, characteristically, in the chapter dealing with the theory of probability for a finite set of events. The strictly infinitary parts of the theory are purely mathematical, and do not correspond to anything in the empirical world. He borrows from [von Mises, 1931] the title, 'the relation to the world of experience' for his Section I.2, and follows von Mises's presentation of the conditions of the applicability of the theory to the world of experience. The application of probability takes place according to the following scheme (p. 3):

1. A certain complex S of unlimitedly repeatable conditions is assumed.
2. One investigates certain events that may appear in the realization of the conditions S. In individual cases of the conditions, the events appear in general in different ways. Let E be the set of possible variants $x_1, x_2, \ldots$ of how the events appear. The set E contains all variants we hold a priori for possible.
3. If the variant appearing after the realization of conditions S belongs to the set A, we say the event A appeared.
4. Under certain conditions, one can assume that to the event A a real number $P(A)$ is attached such that

A. If the conditions S are repeated a great number of times n, one can be *practically certain* that the relative frequency m/n of occurrence of A differs only little from $P(A)$.

B. If $P(A)$ is very small, one can be practically certain that A does not appear in a single realization of the conditions S.

As noted above, Kolmogorov saw probabilistic independence, and weakened but analogous conditions, as the notions that distinguish probability theory from measure theory in general. The application of probability calls for a justification of independence. And indeed, we find him writing: 'After the philosophy of natural science has explained the much debated question concerning the character of the concept of probability itself, one of its most important tasks is the following: to make precise the conditions under which any given real phenomena can be held mutually independent' (pp. 8–9). He adds that 'this question falls outside the scope of our book'. Much later he said he did not answer the problem of application of probability in 1933 because he did not know what the answer should be.

5 INFINITE FIELDS OF PROBABILITY

Kolmogorov's Chapter II is devoted to infinite fields of probability. The two mathematical novelties by which his book differs from previous formulations of measure-theoretic probability, concern such fields. These are the theory of conditional probabilities and the construction of a random process as a probability measure over an infinite-dimensional product space.

The presentation of infinite fields of probability begins with the *axiom of continuity* (p. 13). Let $\bigcap_i A_i$ and $\bigcup_i A_i$ be the finite or denumerable intersections and unions of $A_1, A_2, A_3, \ldots$. Axiom VI reads:

VI. For a descending sequence (1) $A_1 \supset A_2 \supset \cdots$ of events from $\mathscr{F}$ with (2) $\bigcap_i A_i = \emptyset$, it holds that (3) $\lim P(A_i) = 0$ as $i \to \infty$.

If a field of probability is finite, let A_k be the smallest set in (1). Then, since $\bigcap_i A_i = A_1 \cap \cdots \cap A_k = \emptyset$ by (2), $A_k = \emptyset \in \mathscr{F}$. Therefore $P(A_k) = 0$ so (3) follows (pp. 13–14). This also proves that the system is consistent and non-categorical. The continuity axiom is, by an easy argument, equivalent to denumerable additivity (p. 14), or σ-additivity as it is also called. Assume that $A_1, A_2, A_3, \ldots$ form a disjoint sequence of events:

$$P\left(\bigcup_i A_i\right) = \sum_i P(A_i). \tag{4}$$

Denumerable additivity is not a universally accepted property of probability measures. Kolmogorov sees it as a mathematical convention, a view based on his finitism (p. 14):

> Because the new axiom is essential only for infinite fields of probability, it would hardly be possible to explain its empirical meaning in the way sketched

> for axioms I–V in section 2 of the first chapter. In the description of any really observable random process, one can obtain only finite fields of probability. Infinite fields of probability appear only as idealized schemes of real random processes. *We delimit ourselves arbitrarily to schemes which fulfil the continuity axiom VI.*

Axiom VI would not work if the field of events were not closed with respect to denumerable unions and intersections. Kolmogorov refers to Hausdorff's *Mengenlehre* the construction of the smallest σ-field $B\mathscr{F}$ over a given field of sets $\mathscr{F}$. Then he goes on to the extension of a denumerably additive probability P over a field $\mathscr{F}$ into a σ-field $B\mathscr{F}$ (p. 16). This is followed by a remark to the effect that even if the events A from $\mathscr{F}$ can be taken as (possibly only approximately) observable real events, it does not follow that this would be the case for the sets of $B\mathscr{F}$. The extended field of probability $(B\mathscr{F}, P)$ remains a purely mathematical construction (p. 16). As we noted above, the infinitary events from $B\mathscr{F}$ are 'ideal events' in Kolmogorov, and their status is the one that Hilbert gave for ideal elements in mathematics in general. 'If the use of probabilities of these ideal events leads to a determination of the probability of a real event in $\mathscr{F}$, it is obviously automatically acceptable from an empirical point of view'(p. 16).

In Chapter III, random variables are defined as measurable functions. Then the consistency conditions for a system of finite-dimensional distributions are laid down, as pertaining to n random variables (p. 24). These conditions require that the distributions for any k variables, with $k < n$, coincide with marginals of n-dimensional distributions. With the systems of finite-dimensional distributions, all the prerequisites have been laid for the next paragraph III.4, in which the elementary events are points in an infinite-dimensional space. Such product spaces had been considered in measure theory earlier, and even their probabilistic significance had been under some attention. In Kolmogorov's treatise, the product space construction is made for the purpose of a measure-theoretic treatment of stochastic processes as follows:

Let M be any set. Then the probability space to be considered is the set $R^M = \{x_\mu\}$ where $\mu \in M$. To given n indices $\mu_1, \ldots, \mu_n$ there corresponds an n-dimensional subspace R^n. A set A is a *cylinder set* if it is the inverse of the projection from R^M to R^n of a set $A' \subset R^n$, for some n. If A' is a Borel set, A also is by definition. Let $\mathscr{F}^M$ be the Borel sets of R^M thus obtained. Its Borel extension is $B\mathscr{F}^M$. If a probability P is given over $\mathscr{F}^M$, $A \in \mathscr{F}^M$ is a cylinder set and its probability $P(A)$ is obtained as follows. There is a set A' of R^n to which A projects such that $P(A) = P_{\mu_1 \cdots \mu_n}(A')$. Here the latter probability is well determined since it is the probability for the n random variables $x_1, \ldots, x_n$. A system of finite-dimensional distributions determines in this way the probabilities for all Borel sets. Therefore it determines a probability P on $\mathscr{F}^M$. By the extension theorem the same holds for $B\mathscr{F}^M$. Kolmogorov's *Hauptsatz*, or what is often called his extension theorem, now shows that a consistent system of finite-dimensional distributions determines a probability P on $\mathscr{F}^M$ and $B\mathscr{F}^M$ fulfilling the axioms I–VI (p. 27).

The Kolmogorov extension theorem allows for two things: firstly, the discussion of the strong limit theorems of probability in a systematic setting. These theorems typically state that the limit of a denumerable sequence has with probability one a certain property. That probability one is, after Kolmogorov, the same as the measure in an infinite-dimensional

space of all possible sequences. The systematic reason for the connection between strong limit laws and measure-theoretic probability is brought into clear light. Secondly, the extension theorem allows of the construction of a probability law of a stochastic process with an arbitrary index set M, starting from the finite-dimensional distributions. (In keeping with the synthetic mode of presentation, stochastic processes are mentioned only later in the book, on p. 39.) The probability is not defined on *all* subsets of R^M, however, if the index set M is continuous. Kolmogorov's example is the set defined by requiring x_μ to be below a given bound for each μ (p. 26). There are relatively simple-sounding events the probability of which is not well defined by his procedure. Some however think they properly only sound simple, but are not in fact intrinsically well defined [Doob, 1953].

In Chapter V Kolmogorov develops the second of the two essential novelties of his book, the theory of conditional probabilities for infinite sets of elementary events. It applies to cases in which the condition has zero probability, such as one encounters in the theory of Brownian motion for example. Conditional probabilities are defined as random variables, the case of a zero probability condition being handled with what is today called the Radon–Nikodym theorem.

6 THE IMPACT OF THE *GRUNDBEGRIFFE*

Before entering into the immediate reception of the *Grundbegriffe*, we add some remarks on the rest of its contents. The last Chapter VI is a treatment of laws of large numbers to which topic Kolmogorov had contributed continuously, since his first joint paper with A.Y. Khintchine in 1925. An appendix of the book contains a purely infinitistic theorem, namely what is called a zero-one law. As was mentioned, it escapes the concrete sense of chance and probability according to Kolmogorov, whereas some other infinitistic results allow of a finitistic reformulation. The zero-one law says that under rather general conditions, the probability of convergence of a sequence can obtain only the values zero or one. The *Grundbegriffe* ends with a bibliography on previous works on probability of foundational interest.

The mathematical novelties of Kolmogorov's book, besides the organization of the axiomatization, were the construction of stochastic processes and the general theory of conditional expectations, conditional probabilities in particular. He says himself in the preface that 'these new questions arose out of necessity from certain very concrete physical questions'. As we have seen, conditional probabilities where the condition is drawn from a continuous set (thus having in general zero probability), appear at once in the theory of stochastic processes.

We turn now to the reception of the Kolmogorovian measure-theoretic probabilities. The new approach has later certainly been seen as a revolution that made earlier theories obsolete. Some such later descriptions by contemporaries or near contemporaries of Kolmogorov are [Doob, 1989] and [Cramér, 1976]. In Doob one finds bewilderment as to why the approach was 'not immediately universally accepted at once', on the ground of 'the uncontroversial nature of the measure-theoretic approach' [p. 820]. We also read that Kolmogorov's book 'transformed the character of the calculus of probabilities, moving it into mathematics from its previous state as a collection of calculations inspired by a vague

nonmathematical context' [p. 818]. Cramér puts his words more carefully: 'Looking back towards the beginning of a new era in mathematical probability theory, it seems evident that a real breakthrough came with the publication of Kolmogorov's book' [1976, 519]. Already in 1939 in his review of 'Lines of development in the calculus of probability' he had emphasized the continuity, instead of an opposition, between classical and modern probability. He explains briefly the measure-theoretic basis and goes on to show how the classical problems appear as special cases in the new theory. 'Here, too, the case turns out of so many revolutionary ideas: the development does not take place as spontaneously as may seem on a first look. The central new ideas partly are only a consequent, necessary redevelopment of a common property of thoughts that one can follow a long way back in time' [Cramér, 1939, 67].

In his book [Cramér, 1937] a vague frequentist idea of probability is first introduced. Different axiomatizations are always possible, he says. Then he explains a little the frequentist theory of von Mises. Its difficulties, in defining the irregularity of collectives, justify the choice of Kolmogorov's axiomatic measure-theoretic approach, 'at least for the time being' [1937, 4]. Thus, measure-theoretic probability is not seen as any necessity, logical, mathematical, historical, or what have you, and nor was it a novelty of Kolmogorov's.

A long development culminated in Kolmogorov's monograph. It is undeniable that its appearance meant a remarkable advancement in the mathematics of probability. This was mainly felt in the theory of stochastic processes where Kolmogorov's use of infinite product spaces met with immediate approval, but measure-theoretic probability found other uses, too. One of the first to join Kolmogorov was Khintchine, who was at the time developing a probabilistic approach to ergodic theory and the theory of stationary processes. Eberhard Hopf had been using measure theory in his studies of dynamical systems. Hopf in his great paper on probabilistic aspects of dynamical systems [1934], immediately took advantage of Kolmogorov's measure-theoretic probabilities. J.L. Doob started at once after 1933 to develop the theory of stochastic processes. His systematic papers from the late 1930s are devoted to the study of discrete and continuous-time stochastic processes. Measure-theoretic probabilities were also taken into use by Cramér and Lévy in their books, which both appeared in 1937. Even on the basis of this very partial list, one can conclude that many of the leading researchers in mathematical probability soon absorbed Kolmogorov's measure-theoretic probabilities. Their 'universal acceptance', on the other hand, took its time. This was partly due to resistance from other, competing approaches to probability, notably the theory of von Mises. Bruno de Finetti also systematically refused to think that a measure-theoretic scheme would be more than a useful way of finding examples. Instead he offered an alternative, stemming from his thought that probability theory must have a form immediately appealing to the 'everyday sense' of probability. Anyway, as concerns the reception of measure-theoretic probabilities, it is a fact that textbook expositions did not start appearing until after the Second World War, with the exception of [Cramér, 1937]. His 1946 book *Mathematical methods of statistics* makes systematic use of measure theory, as does [Doob, 1953]. Paul Halmos's book of 1950, *Measure theory*, the standard treatise on its topic for a long time, devotes one chapter to probability measures.

BIBLIOGRAPHY

Cramér, H. 1937. *Random variables and probability distributions*, Cambridge: Cambridge University Press.

Cramér, H. 1939. 'Entwicklungslinien der Wahrscheinlichkeitsrechnung', in *Neuvième congrès des mathématiciens scandinaves*, Helsingfors: Mercators Tryckeri, 67–86.

Cramér, H. 1976. 'Fifty years of probability theory: some personal recollections', *The annals of probability*, *4*, 509–546.

Doob, J.L. 1953. *Stochastic processes*, New York: Wiley.

Doob, J.L. 1989. 'Kolmogorov's early work on convergence theory and foundations', *The annals of probability*, *17*, 815–821.

Hausdorff, F. *Mengenlehre*, Berlin and Leipzig: de Gruyter, 1927. [Repr. New York: Chelsea, 1950.]

Hopf, E. 1934. 'On causality, statistics and probability', *Journal of mathematics and physics* (MIT), *13*, 51–102.

Kolmogorov, A. 1923. 'Une série de Fourier–Lebesgue divergente presque partout', *Fundamenta mathematicae*, *4*, 324–328.

Kolmogorov, A. 1925. 'On the principle of excluded middle' [in Russian], *Matematicheski Sbornik*, *32*, 646–667. [English trans. in J. van Heijenoort (ed.), *From Frege to Gödel*, Cambridge, MA: Harvard University Press, 1967, 414–437.]

Kolmogorov, A. 1926. 'Une série de Fourier–Lebesgue divergente partout', *Comptes rendus de l'Académie des Sciences*, *183*, 1327–1328.

Kolmogorov, A. 1928. 'On operations on sets' [in Russian], *Matematicheski Sbornik*, *35*, 414–422.

Kolmogorov, A. 1929. 'General theory of measure and the calculus of probability' [in Russian], repr. in *Probability theory and mathematical statistics* [in Russian], Moscow: Nauka, 1986, 48–58. [English trans. in *Selected works*, vol. 2, Dordrecht: Kluwer 1992.]

Kolmogorov, A. 1931. 'Ueber die analytischen Methoden in der Wahrscheinlichkeitsrechnung', *Mathematische Annalen*, *104*, 415–458.

von Mises, R. 1931. *Wahrscheinlichkeitsrechnung und ihre Anwendung in der Statistik und theoretischen Physik*, Leipzig and Vienna: Deuticke.

von Plato, J. 1994. *Creating modern probability*, Cambridge: Cambridge University Press.

CHAPTER 76

H. SEIFERT AND W. THRELFALL (1934) AND P.S. ALEXANDROFF AND H. HOPF (1935), BOOKS ON TOPOLOGY

Alain Herreman

In the early 20th century algebraic topology was a discipline at once young and in full elaboration. These two books, taking part in the development of modern algebra and set theory, succeeded in offering for the first time a clear and coherent presentation of this already vast discipline, each book following different points of view. For this reason they were immediately accepted and use as reference works for quite some time.

Seifert and Threlfall

First publication. *Lehrbuch der Topologie*, Leipzig: Teubner, 1934. vii + 353 pages.

English translation. *A textbook of topology* (trans. Michael A. Goldman), New York: Academic Press, 1980. [Includes also Seifert, *Topology of 3-dimensional fibered spaces* (trans. W. Heil, ed. J.S. Birman and J. Eisner).]

Alexandroff and Hopf

First publication. *Topologie*, Berlin: J. Springer, 1935 (*Die Grundlehren der mathematischen Wissenschaften*, volume 45). xiv + 636 pages.

Photoreprints. New York: Chelsea, 1965, 1972. Also Berlin and New York: Springer, 1974.

Related articles: Cantor (§46), Riemann on geometry (§39), Urysohn and Menger (§60), van der Waerden (§70)

1 ALGEBRAIC TOPOLOGY PRIOR TO 1934

1.1 Poincaré's memoirs and combinatorial analysis situs

The books of Seifert and Threlfall and Alexandroff and Hopf deal with the subject known as 'algebraic topology'. From their titles however, they are simply books on 'topology'.

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

Indeed, it was not until 1942, with the publication of the book *Algebraic topology* by Solomon Lefschetz (1884–1972), that the adjective 'algebraic' was coupled with 'topology' in the title of a book or article. Nevertheless, as early as 1932, Alexandroff indicated in a note on a little book [Alexandroff, 1932] that he preferred this name to 'combinatorial topology', 'since we consider much broader applications of algebraic methods and concepts than the word "combinatorial" would include' [Alexandroff, 1961, 27]. In addition to the word 'Topology', already used by J. Listing (1808–1882) in [Listing, 1847], the expression 'Analysis situs', introduced by G.W. Leibniz (1646–1716), was also in common use up to the beginning of the 1930s, to designate that branch of mathematics defined by Bernhard Riemann (1826–1866) as 'the study of continuous magnitudes where one does not consider them as existing independently of their position or as measurable in terms of each other, but where one studies only the relative situation of places and regions, entirely without reference to any metric relation' ([Riemann, 1857]: compare §39).

'Analysis situs' was also the designation adopted by Henri Poincaré (1854–1912) in his series of six memoirs on the subject [Poincaré, 1895–1904]. It is in these 'fascinating and exasperating' memoirs [Dieudonné, 1989] that Poincaré introduces most of the basic notions, methods, theorems and conjectures relating to homology and the fundamental group. In the next 30 years, most of the papers inspired by these memoirs attempted to give a more satisfying presentation of certain parts by means of more general or more suitable definitions. Thus there developed a 'combinatorial analysis situs' or 'combinatorial topology', which covered a great variety of approaches. Among them the spaces considered (called today 'manifolds', 'complexes', 'chains', and so on) composed of cells (also called today 'simplexes'), like a polyhedron is composed of faces. All these designations ('topology', 'analysis situs', 'combinatorial analysis situs', 'algebraic topology') reflect only partially the diversity of the definitions adopted: two articles then rarely had the same definition of manifolds [Herreman, 2000].

1.2 Books on topology published before 1934–1935

It was only in 1922 that the first book on the subject appeared: *Analysis situs*, by Oswald Veblen (1880–1960) [Veblen, 1922]. This elementary book of 194 pages was the main reference for nearly 10 years and was re-issued in 1931. It proposed a presentation that attempted to reconcile the arithmetical approach (via matrices and numbers) with the geometric in the study of homology with coefficients in integers or reduced modulo 2. Its geometric nature is clear from the way it treats spaces in order of dimension: first spaces of dimension 1 (linear graphs), then complexes and manifolds of dimension 2, and finally complexes of dimension n. It was from this book that Saunders MacLane tried to learn the subject: 'from such a book, without a teacher, it was exceedingly difficult to understand combinatorial topology; in 1931 I tried with Veblen's book and failed' [MacLane, 1986, 306].

Later there appeared a book by a colleague of Veblen, namely Lefschetz. Containing more than 400 pages, [Lefschetz, 1930] includes duality theorems, the theory of intersection of chains, intersection theory and the theory of fixed points of continuous maps between two manifolds, with applications to analytic and algebraic manifolds. Using the theory of sets, still very geometric in this book, the author was able to extend the notion of

homology to that of homology relative to a subset of a manifold, and to embrace in a single theorem the duality theorems of Poincaré and J.W. Alexander (1888–1971), on whom see section 3.1. It covered the subject better than its predecessor, but was still judged 'quite difficult to read' by the young Hassler Whitney (1907–1989) [Whitney, 1942].

Among available books devoted to point set topology one may mention [Young and Young, 1906; Hausdorff, 1914; Fréchet, 1928] and [Kuratowski, 1933].

2 SEIFERT AND THRELFALL, *LEHRBUCH DER TOPOLOGIE* (1934)

2.1 Biographical notes on Seifert and Threlfall

Herbert Seifert (1907–1996) began his study of mathematics and physics at the Technical University of Dresden in 1926. In the following year, he followed for the first time the topology courses of William Threlfall (1888–1949). This encounter quickly developed into a productive friendship which led to numerous joint publications, including two books: [Seifert and Threlfall, 1938] preceded by the book under discussion, which in part came out of these courses.

Seifert spent part of the academic year 1928–1929 at Göttingen, where he met Alexandroff and Hopf (see section 3.1 below). He obtained his doctorate in 1930 with a paper on 'Construction of three-dimensional closed spaces' [Seifert, 1931]. He was appointed professor at the University of Heidelberg in 1935 following the dismissal of Heinrich Liebmann by the Nazis. He occupied this position until his retirement, except for the war years, when he worked at the *Institut für Gasdynamik.*

In 1938 Threlfall was appointed professor at the University of Frankfurt-am-Main, where he succeeded C.L. Siegel, who had emigrated to the United States. In 1946 he obtained for his friend a professional position at his university, but the period of separation and the death of Threlfall in 1949 precluded any further collaboration between them [Puppe, 1999].

2.2 The book

The contents of their book is summarised in Table 1. Bearing in mind that 'topology is intimately associated with the theory of groups' (p. 305), the authors devote a chapter to the elements of the theory of groups (defined in terms of generators and relations) that are needed throughout the book. The most general spaces are 'neighbourhood spaces', that is sets of points to which are associated neighbourhoods satisfying the axioms ad hoc. This level of generality serves mainly to define continuous functions, for the book is devoted to the study of less general spaces, namely, complexes. A (simplicial) complex is a 'neighbourhood space which can be simplicially decomposed' (p. 43). The (simplicial) homology groups of a complex are defined using simplicial chains, that is, linear combinations with integer coefficients or reduced modulo 2 of the oriented simplexes of the complex: the homology groups are the 'residue classes of the lattice G^k of closed k-chains relative to the sublattice N^k of null-homologous k-chains' (p. 66).

The book then describes the classical results of combinatorial topology obtained using incidence matrices: the reduction to normal form, torsion numbers, the Euler–Poincaré

Table 1. Contents by chapters of the book by Seifert and Threlfall.

Page	Topics
I: Illustrative material.	
1	The principal problem of topology. Closed surfaces. Isotopy, homotopy, homology. Higher dimensional manifolds.
II: Simplicial complexes.	
22	Neighborhood spaces. Mappings. Point sets in Euclidean spaces. Identification spaces; n-simplexes. Simplicial complexes. The schema of a simplicial complex. Finite, pure, homogeneous complexes. Normal subdivision. Examples of complexes.
III: Homology groups.	
60	Chains. Boundary, closed chains. Homologous chains. Homology groups. Computation of the homology groups in simple cases. Homologies with division. Computation of homology groups from the incidence matrices. Block chains; chains mod 2, connectivity numbers. Euler's formula. Pseudomanifolds and orientability.
IV: Simplicial approximation.	
95	Singular simplexes; singular chains. Singular homology groups. The approximation theorem. Invariance of simplicial homology groups. Prisms in Euclidean spaces. Proof of the approximation theorem. Deformation and simplicial approximation of mappings.
V: Local properties.	
123	Homology groups of a complex at a point. Invariance of dimension. Invariance of the purity of a complex; of its boundary; of pseudomanifolds and of orientability.
VI: Surface topology.	
134	Closed surfaces. Transformation to normal form. Types of normal form: the principal theorem. Surfaces with boundary. Homology groups of surfaces.
VII: The fundamental group.	
154	The fundamental group. Examples: the edge path group of a simplicial complex, and of a surface complex. Generators and relations; edge complexes and closed surfaces. The fundamental and homology groups. Free deformation of closed paths. Fundamental group and deformation of mappings. This group at a point, and of a composite complex.
VIII: Covering complexes.	
188	Unbranched covering complexes. Base path and covering path. Coverings and subgroups of the fundamental group. Universal coverings. Regular coverings. The monodromy group.
IX: 3-dimensional manifolds.	
211	General Principles. Representation by a polyhedron; homology groups; the fundamental group. The Heegaard diagram. 3-dimensional manifolds with boundary; construction of 3-dimensional manifolds out of a knot.

Table 1. (*Continued*)

Page	Topics
X: n-Dimensional manifolds.	
235	Star complexes cell complexes. manifolds; the Poincaré duality theorem. Intersection number of cell chains; dual bases. Cellular approximations. Intersection numbers of singular chains. Invariance of intersection numbers. Examples: orientability and two-sidedness; linking numbers.
XI: Continuous mapping.	
294	The degree of a mapping. a trace formula. A fixed point formula. Applications.
XII: Auxiliary theorems from the theory of groups.	
305	Generators and relations. Homomorphic mapping and factor groups. Abelianization of groups. Free and direct products. Abelian groups. The normal form of integer matrices.
328	Comments.
341	Bibliography. [End 353.]

characteristic, and so on. Topological invariance is proved using singular homology groups defined in terms of singular chains, that is, integer linear combinations of singular oriented non-degenerate simplexes. The latter are topological invariants by definition, and the topological invariance of simplicial homology groups then follows from a deformation theorem, which reduces homologies between singular chains to homologies between simplicial chains of a sufficiently fine subdivision of the complex. This is a clearer exposition of the proof given by Veblen [1922], which is in turn taken from the paper [Alexander, 1915] and is essentially contained in Poincaré's memoirs.

The authors introduce 'homology groups at a point': the homology groups of the complex formed by the simplexes not containing that point. These are also topological invariants, and they used this fact to prove the topological invariance of dimension. We recall that the definition of dimension of a space and the proof of its invariance had been classical problems since the work of Georg Cantor (1845–1918): several definitions of dimension and various proofs of invariance had been given (§66), the first valid one by L.E.J. Brouwer (1881–1966) in [Brouwer, 1911].

Having introduced these elements of combinatorial topology and illustrated them by numerous examples, the authors show that the Euler characteristic together with orientability are sufficient to characterize a closed surface defined by identifying the sides of polygons. Even if this result and proof were not entirely new, their presentation remains a model. Having in mind the problem of classifying complexes of dimension greater than 2, the authors introduce the fundamental group, defined as the group of homotopy classes of closed paths in a complex. They make a connection between this group and the 'first homology group by proving that the first homology group of a connected complex is the Abelianised fundamental group'. They go on to show that the fundamental group is also the 'group of covering transformations of the universal covering complex'. These notices are then ap-

plied to manifolds of dimension 3, but without solving completely the problem of their classification.

The authors then restrict themselves to manifolds, which they define as 'connected, finite, n-dimensional complexes which at every point have the same homology groups as the $(n-1)$-sphere'. This definition, which consists of retaining only the homological characterization of simplicial neighbourhoods, had already been considered by various mathematicians, notably by Alexander and E. van Kampen, and also in [Lefschetz, 1930]. The property of being a manifold in this sense is thus topologically invariant, which is not the case with other definitions. This definition also makes it possible to state and prove the Poincaré duality theorem, either starting from incidence matrices or by constructing dual bases using intersection numbers. Both proofs are given. Finally, one chapter is devoted to the theory of the degree of a map, developed by Brouwer, applied to fixed-point theorems and to the 'trace formula' of Hopf.

3 ALEXANDROFF AND HOPF, *TOPOLOGIE* (1935)

3.1 Biographical notes on Alexandroff and Hopf

P.S. Alexandroff (1896–1982) was a student of Nikolai Nikolaevich Luzin (1883–1950), and his first papers were devoted to general topology. In 1923, at the age of 27, he and his friend P.S. Urysohn (1898–1924) went to Göttingen, which was then one of the main centres of mathematics. There they met, among others, Emmy Noether (1882–1935) and David Hilbert (1862–1943). During the summer of 1924, they went to Holland together to meet Brouwer. After the death of his friend in a bathing accident in Britanny (§66.3) Alexandroff returned to work with Brouwer for part of the years 1925 and 1926.

It was in the course of his regular visits to Göttingen between then and 1932 that Alexandroff struck up a friendship with Heinz Hopf (1894–1971), who was two years his senior. Hopf had already followed, at Breslau and then Berlin, the courses of Erhard Schmidt (1876–1959) on the theory of sets and the work of Brouwer. He first came to Göttingen in 1925. It was in December of that year, at a dinner at Brouwer's house, that Noether suggested replacing the Betti numbers by groups. Hopf used this idea to good effect in generalizing the Euler–Poincaré formula [Hopf, 1928].

Supported by a Rockefeller grant, Alexandroff and Hopf went to Princeton together, where they spent the academic year 1927–1928. They attended the courses of, and held discussions with, those of the most eminent topologists of the time: Veblen, Lefschetz and Alexander. The last-named is best known for his two proofs of the invariance of Betti numbers and connection numbers (Betti numbers reduced modulo 2) [Alexander, 1915], his counterexamples showing that the Betti numbers and the fundamental group are insufficient to classify manifolds of dimension 3, and the duality theorem that bears his name today [Alexander, 1922].

Alexandroff was appointed Professor of Mathematics at the State University of Moscow in 1929. In 1931, Hopf was appointed Professor of Mathematics at the Technical High School of Zürich, where he remained until his retirement in 1965 [Arboleda, 1979; Frei and Stammbach, 1999].

3.2 The book

The book was written for the Springer series *Grundlagen der mathematischen Wissenschaften* ('Foundations of the mathematical sciences') on the initiative of Richard Courant, who had suggested the project to the future authors in 1928. Alexandroff being in Moscow and Hopf in Zürich, their collaboration was carried out chiefly by letter. The International Congress in Zürich (5–12 September 1932), of which Hopf was one of the organizers, enabled the two authors to meet; and it was at the first international conference on topology (4–10 September 1935), this time organized by Alexandroff in Moscow, that they finished the book. Its contents are summarised in Table 2.

The object of the book is not to present 'the whole of topology' but 'topology as a whole', which is perhaps more ambitious. Instead of opting for general topology or combinatorial topology, it combines these two approaches in the way that Brouwer, to whom the book is dedicated, had been the first to suggest. The first part is devoted to the elementary notions of general topology: open and closed sets, metric and topological spaces (defined in terms of closure), continuous maps, separation axioms, and compact, bicompact and complete spaces.

Table 2. Contents by Part of the book by Alexandroff and Hopf.

Page	**Topics**
First Part: Basic concepts of set-theoretic topology.	
24	Topological and metric spaces.
83	Compact spaces.
Second Part: Topology of complexes.	
124	Polyhedra and its cell decomposition [*Zellenzerlegungen*].
154	Vertices and coefficient domains.
205	Betti groups.
240	Sub-division and decomposition of complexes.
273	Special questions from the theory of complexes.
Third Part: Topological invariance theorems and related development of concepts.	
313	Simplicial approximations of continuous mappings.
347	Canonical displacement [*Verschiebungen*]. Again on the invariance of dimension number and of Betti groups. General concept of dimension.
379	The decomposition theorem for Euclidean space. Further invariance theorems.
Fourth Part: Linking in Euclidean space. Continuous mappings of polyhedra.	
409	Linking theory. Alexander's duality theorem.
457	Brouwer's mapping degree. Kronecker's characteristic.
498	Homotopy und extension theorems for mappings.
527	Fixed points.
554	*Appendix I:* Abelian groups.
594	*Appendix II:* R^n and its convex cells.
617	List of topological books.
618	Bibliography.
622	Subject index. [End 636.]

The second part introduces the basic concepts of combinatorial topology: 'curved' polyhedra and their subdivisions. Polyhedra are defined using complexes, where a complex is a finite or countable set of simplexes (line segments, triangles, tetrahedral, and so on) with ad hoc conditions on the relations of incidence between them. A polyhedra is then the set of points belonging to at least one cell of the complex. The need to distinguish clearly between the set of cells (the complex) and the set of points (the polyhedron) was emphasized for the first time in [Alexandroff, 1932], when confusion was still rife [Herreman, 2000]. A curved polyhedron ('krumme Polyeder') is then a topological space homeomorphic to a polyhedron. It is thus a geometrical object.

Algebra, in the form of the theory of abelian groups, enters the picture via algebraic complexes. An algebraic complex associates a set of vertices and an abelian group: the vertices suffice to define the simplexes of an oriented complex, and an *algebraic complex* is then a linear combination $C = \sum t^i x_i$, where the x_i are simplexes and the t^i elements of the abelian group or of a ring: the integers, integers modulo m, rational numbers, or rational numbers modulo 1 in which L. Pontrjagin (1908–1988) had shown an interest. These complexes enable one to define the Betti groups as quotients of the group of cycles with coefficients in a group by the group of cycles that are boundaries. The authors make precise the influence of the choice of coefficient group on the Betti groups and establish the invariance of the latter by subdivision of the polyhedron.

The third part of the book is devoted to polyhedra and to proving the topological invariance of the dimension and the Betti groups. Several proofs are given: one makes use of the link established by simplicial approximation between homotopy classes and homology classes, and another involves the nerve of a covering of a metric space, a notion introduced by Alexandroff and inspired by the dual polyhedra of Poincaré and the simplicial techniques adopted by Brouwer.

The fourth part is devoted to the theory of intersection and linking numbers in Euclidean space of dimension n that enables one to prove the Alexander duality theorem. One chapter is given over to the Lefschetz fixed-point formula for a transformation of a polyhedron onto itself.

4 RECEPTION OF THE TWO BOOKS

The recognition of a new branch of mathematics, analysis situs, can be attributed to Leibniz. It had a name and a well-established and stable definition; but hardly any definitions, theorems or proofs were shared by many of the numerous contributions to this area in the second half of the 19th century [Pont, 1974]. Neither did any basic text serve as a common reference. This changed with the appearance of Poincaré's memoirs: they introduce a set of definitions, theorems, proofs and methods that were to be quickly taken up by the international mathematical community. From then on, analysis situs had its reference texts.

Forty years later, the book of Seifert and Threlfall marked a new era in this history. The elements of homology theory it contains (simplicial homology groups, incidence matrices, a proof of the invariance of singular homology, the theory of manifolds, Poincaré's duality

theorem, intersection and linking, and fixed-point theorems) differ little from those presented four years sooner in [Lefschetz, 1930]. The organization of the body of the theory then seems to have become pretty well fixed: the results and methods of this theory became 'standard'. But Seifert and Threlfall present the material with unprecedented and enduring clarity. This book, while it contains no truly original results, gives the first clear exposition of ideas, which makes it accessible to students. It is this that makes it an important event in the history of algebraic topology. It gives the reader access to common seemingly well-established theories. It is moreover probably more a consequence or simply an effect of the book than the state of a discipline then in full revitalisation.

Whatever be the case, the reading of it allowed economies to be made in the reading of other texts, Poincaré's difficult memoirs in particular. It thus modifies the rapport of the reader with earlier publications. This is another effect of its clarity and status as a reference manual. Its numerous historical notes forge new links with texts that will from then on be read less or even not at all.

The 'Julia seminar' testified to the reception of both books. Held in Paris since 1933, this seminar was led by a group of young mathematicians: J. Dieudonné, A. Weil, C. Chevalley, J. Leray, R. de Possel, C. Ehresmann, P. Dubreil and F. Marty. Among them one recognises some of the mathematicians who founded Bourbaki two years later, after which Weil, Dubreil and Chevalley moved to Göttingen. The theme of the seminar in its first year was the 'theory of groups and algebras'. It provided a forum for the study of the works of G. Frobenius, Noether, E. Artin, A. Speiser, H. Hasse and B. L. van der Waerden, the first volume of whose *Moderne Algebra* had appeared in 1930 (§70). The second year was devoted to Hilbert spaces, and the following year, 1935–1936, to topology, where our two books played their part. André Weil, who had met Alexandroff at Göttingen in 1927, chose Alexandroff and Hopf as the reference book for his accounts of the 'applications of homological invariants to the characterization of classes of representation' and 'intersection numbers and topological degrees'. It was also the reference for de Possel's account of 'fixed points of transformation'. Seifert and Threlfall served similarly for F. Marty's account of 'coverings—the fundamental group'. Poincaré's memoirs were never cited by the young French mathematicians, and neither was Veblen's book. Lefschetz's book was only mentioned by Ehresmann, Chevalley and Weil. Ten years later, H. Cartan again referred 'for all fundamental notions' to Alexandroff and Hopf [Cartan, 1945, Note, p. 2], and J. Leray mentioned the first two chapters of 'the excellent treatise of Messrs. Alexandroff and Hopf' [Leray, 1945].

It was the same story in the United States. Thus Whitney, who was of the same age as the young French mathematicians, wrote [Whitney, 1942]:

> In 1934, the appearance of [Seifert and Threlfall] gave the young student an excellent first text in the field. In one subject, the fundamental group (with applications), it fills a large gap in earlier works. Soon after (1935) [Alexandroff and Hopf] was published. It is a very full treatment [...]. Both [Seifert and Threlfall] and [Alexandroff and Hopf] are written in a clear, detailed style.

Peter Hilton spoke similarly of Alexandroff and Hopf: 'This was an extremely influential book and was a sort of bible for the study of algebraic topology. It was a very beautifully written work' [Hilton, 1988, 286].

Each of these books refer to subsequent volumes in their introductions: Seifert and Threlfall intended to treat Alexander duality in a second volume, and Alexandroff and Hopf were even more precise in specifying two additional books, the first to be devoted to topological spaces in full generality, and the second to varieties and the fundamental group. But none of these saw the light of day. At the time when these two books appeared, two discoveries were announced at the Moscow Congress that transformed algebraic topology: cohomology was discovered independently by Alexander and A.N. Kolmogorov (1903–1987), and homotopy groups were defined in all dimensions by W. Hurewicz (1904–1956).

Political events also had an effect on the geography of the mathematical world. At the end of the 1920s and beginning of the 1930s, mathematicians came from all over the world to Göttingen: Garrett Birkhoff, Whitney and Mac Lane, to name but a few. The advent of Naziism emptied the University of Göttingen and transplanted the heart of mathematics in the United States, especially in Princeton, where a new algebraization of topology took place. Partly as a result of these developments and this algebraization, our two books were quickly overtaken by the specialists. To add one or two volumes would no longer suffice; a rewriting would have been necessary. It was this that led to Lefschetz in 1942 to publish his book *Algebraic topology*, which bore the stamp of this new generation of mathematicians at Princeton, such as Mac Lane, S. Eilenberg, N.E. Steenrod, Whitney and Chevalley.

From that time onwards, our two books could no longer be regarded as more than introductory texts. Seifert and Threlfall seems an obvious candidate for this role: Alexandroff and Hopf is not properly speaking a textbook, but rather a coherent and systematic exposition of a large part of the subject. It gains its generality and coherence largely from the set-theoretic language in which it is uncompromisingly written: a good number of its distinctions and developments are bound up with this choice. Seifert and Threlfall lies closer to basic geometric intuition: this does not lose its interest or its value in the later evolution of the theory, which needs to reflect that intuition. Moreover, some theorems in topology are valid, or known to be valid, only in sufficiently high dimensions: a distinction appeared between topology in general and topology in low dimensions. The need has become increasingly obvious for a specific study of low-dimensional topology, as opposed to investigations of the widest generality. It turns out that Alexandroff and Hopf fits into the latter category, while Seifert and Threlfall can serve as an introduction to the former.

BIBLIOGRAPHY

Alexander, J.W. 1915. 'A proof of the invariance of certain constants in Analysis Situs', *Transactions of the American Mathematical Society*, *16*, 148–154.

Alexander, J.W. 1922. 'A proof and extension of the Jordan–Brouwer separation theorem', *Transactions of the American Mathematical Society*, *23*, 333–349.

Alexandroff, P.S. 1932. *Einfachste Grundbegriffe der Topologie*, Berlin: Springer Verlag. [English trans.: *Elementary concepts of topology* (trans. A.E. Farley), New York: Dover, 1961.]

Arboleda, L.C. 1979. 'Les débuts de l'école topologique soviétique: notes sur les lettres de Paul S. Alexandroff et Paul Urysohn à Maurice Fréchet', *Archives for history of exact sciences*, *20*, 73–89.

Brouwer, L.E.J. 1911. 'Beweis der Invarianz der Dimensionzahl', *Mathematische Annalen*, *70*, 161–165. [Repr. in *Collected works*, vol. 2, 430–435.]

Cartan, H. 1945. 'Méthodes modernes en topologie algébrique', *Commentarii Mathematicae Helveticii*, *18*, 1–15.
Dieudonné, J. 1989. *A history of algebraic and differential topology 1900–1960*, Boston: Birkhäuser.
Fréchet, M. 1928. *Les espaces abstraits*, Paris: Gauthier–Villars.
Frei, G. and Stammbach, U. 1999. 'Heinz Hopf', in [James, 1999], 991–1008.
Hausdorff, F. 1914. *Grundzüge der Mengenlehre*, Leipzig: Veit. [Repr. New York: Chelsea, 1949.]
Herreman, A. 2000. *La topologie et ses signes. Eléments pour une histoire sémiotique des mathématiques*, Paris: L'Harmattan.
Hilton, P. 1988. 'A brief and subjective history of homology and homotopy theory in this century', *Mathematics magazine*, *6*, 282–291.
Hopf, H. 1928. 'Eine Verallgemeinerung der Euler–Poincaré Formel', *Nachrichten der Gesellschaft der Wissenschaften zu Göttingen*, 127–136.
James, I.M. (ed.) 1999. *History of topology*, Amsterdam: North-Holland.
Kuratowski, C. 1933. *Topologie*, vol. 1, Warsaw and Lwow: PWN.
Lefschetz, S. 1930. *Topology*, New York: American Mathematical Society.
Lefschetz, S. 1942. *Algebraic topology*, Providence: American Mathematical Society.
Leray, J. 1945. 'Sur la forme des espaces topologiques et sur les points fixes des représentations', *Journal de mathématiques pures et appliquées*, *24*, 95–167.
Listing, H.B. 1847. 'Vorstudien zur Topologie', *Göttinger Studien*, 811–875.
MacLane, S. 1986. 'Topology becomes algebraic with Vietoris and Noether', *Journal of pure and applied algebra*, *39*, 305–307.
Poincaré, H. 1895. 'Analysis Situs', *Journal de l'Ecole Polytechnique*, *(2) 1*, 1–123. [This and the later papers repr. in *Œuvres*, vol. 6 (1953).]
Poincaré, H. 1899. 'Complément à l'analysis situs', *Rendiconti del Circolo Matematico di Palermo*, *13*, 285–343.
Poincaré, H. 1900. 'Second complément à l'Analysis Situs', *Proceedings of the London Mathematical Society*, *32*, 277–308.
Poincaré, H. 1902. 'Sur les cycles des surfaces algébriques; Quatrième complément à l'Analysis Situs', *Journal de mathématiques pures et appliquées*, *(5) 8*, 169–214.
Poincaré, H. 1904. 'Cinquième complément à l'Analysis Situs', *Rendiconti del Circolo Matematico di Palermo*, *18*, 45–110.
Pont, J.-C. 1974. *La topologie algèbrique des origines à Poincaré*, Paris: PUF.
Puppe, D. 1999. 'Herbert Seifert', in [James, 1999], 1021–1027.
Riemann, B. 1857. 'Theorie der Abel'schen Functionen', *Journal für die reine und angewandte Mathematik*, *54*, 101–155.
Seifert, H. 1931. 'Konstruktion dreidimensionaler geschlossener Räume', *Berichte der säschsichen Akademie der Wissenschaften zu Leipzig*, *83*, 26–66.
Seifert, H. and Threlfall, W. 1938. *Variationsrechnung im Grossen*, Leipzig: Teubner.
Veblen, O. 1922. *Analysis situs*, New York: American Mathematical Society (Colloquium Publications). [Repr. 1931.]
Whitney, H. 1942. Review of [Hopf, 1941], *Mathematical reviews*, *3*, 316.
Young, W.H. and Young, G.C. 1906. *The theory of sets of points*, Cambridge: Cambridge University Press. [Repr. New York: Chelsea, 1972.]

CHAPTER 77

DAVID HILBERT AND PAUL BERNAYS, *GRUNDLAGEN DER MATHEMATIK*, FIRST EDITION (1934, 1939)

Wilfried Sieg and Mark Ravaglia

In these two volumes, Hilbert and Bernays present systematically their proof-theoretic investigations and a wide range of current results, such as Herbrand's theorems and Gödel's incompleteness theorems. The second volume has a number of supplements, in which they discuss some specialized topics, for example, the development of mathematical analysis and the unsolvability of the decision problem.

First publication. 2 volumes, Berlin: Verlag Julius Springer, 1934, 1939 (*Die Grundlehren der Mathematischen Wissenschaften*, vols. 40 and 50). 479 + 506 pages.

Second edition. 2 volumes, same publisher, 1968–1970. 472 + 561 pages. [Revisions detailed in forewords written by Bernays.]

French translation of the second edition. Fondements des mathématiques (trans. F. Gaillard, E. Gaillard and M. Guillaume), 2 volumes, Paris: l'Harmattan, 2001.

Russian translation of the second edition. Osnovaniya matematiki (trans. N.M. Nagornyi, ed. S.I. Adyan) 2 vols., Moscow: Nauka Publishing House, 1979. [Repr. 1982.]

Related articles: Dedekind (§43), Dedekind and Peano (§47), Hilbert on geometry (§55), Whitehead and Russell (§61), Gödel (§71).

1 BACKGROUND

The two volumes of *Grundlagen der Mathematik* by David Hilbert (1862–1932) and Paul Bernays (1888–1977) are very special milestones in the development of modern mathematical logic. They were at the forefront of contemporaneous research and presented then current metamathematical results: from consistency proofs (Hilbert and Bernays had obtained in weaker forms during the 1920s) through theorems of Jacques Herbrand (1908–1931)

Landmark Writings in Western Mathematics, 1640–1940
I. Grattan-Guinness (Editor)

and Kurt Gödel (1906–1978) to a sketch of a consistency proof for number theory found by Gerhard Gentzen (1909–1945). This material is supplemented in the second volume by a series of important appendices concerning focused topics, for example, a very elegant formal development of analysis and an incisive presentation of the undecidability of the decision problem. Indeed, the two volumes constitute an encyclopedic synthesis of metamathematical work from the preceding two decades. What is most remarkable, however, is the sheer intellectual force that structures the books: they are penetrating and systematic studies concerned with the foundations of modern mathematics as it emerged in the second half of the 19th century. That emergence was deeply influenced by C.F. Gauss, J.P.G. Dirichlet, Bernhard Riemann, and above all by Richard Dedekind (1831–1916).

Dedekind formulated abstract axiomatic theories within a general logicist framework that was articulated most explicitly in *Was sind und was sollen die Zahlen?* [1888]. His way of formulating theories was used by Hilbert in *Die Grundlagen der Geometrie* [1899] and the paper 'Über den Zahlbegriff' [1900]. Hilbert recognized, as Dedekind had done, the centrality of the consistency problem for such theories. For Dedekind this was a semantic issue, and he tried to resolve it by defining suitable models within logic. However, problematic aspects of Dedekind's broad logicist framework were noticed early by Georg Cantor (1845–1918) and formulated in letters to Hilbert in 1897. Hilbert reformulated the consistency problem as a quasi-syntactic one for his axiomatization of the arithmetic of real numbers, both in his papers (1900) and (1901), in the latter in the second of his Paris problems (§57.2). He demanded that a 'direct proof' be given to establish that no contradiction can be obtained from the axioms in a 'finite number of logical steps'. The point of such a proof was to establish the existence of a 'consistent multiplicity', i.e. a set, satisfying the axioms. At the time, Hilbert thought that a consistency proof could be given 'by means of a careful study and suitable modification of the known methods of reasoning in the theory of irrational numbers'.

Hilbert believed, it seems, that the genetic build-up of the real numbers could be exploited to yield the blueprint for a consistency proof in Dedekind's logicist style. That is supported by Hilbert's treatment of arithmetic in other lectures from that period, but also by a more programmatic statement from the introduction to the notes for his lectures 'Elemente der Euklidischen Geometrie' (summer semester 1899). He maintains there: 'It is important to fix precisely the starting-point of our investigations: as given we consider the laws of pure logic and in particular all of arithmetic' [Toepell, 1986, 203–204]. Hilbert adds parenthetically, 'On the relation between logic and arithmetic cf. Dedekind, *Was sind und was sollen die Zahlen?*'. And, clearly, for Dedekind arithmetic is part of logic.

2 NAÏVE PROOF THEORY

In Dedekind's as well as in Hilbert's systematic developments only the mathematical parts are characterized axiomatically; logic is not given a principled formulation. That changes in 1904 with Hilbert's programmatic call for a simultaneous development of logic and mathematics. However, it is only more than a decade later that an appropriate logical frame is obtained through the careful study of *Principia mathematica* (1910–1913) by A.N. Whitehead (1862–1947) and Bertrand Russell (1872–1970). This fully formal framework is then

recognized as an object of metamathematical investigation—to address the issues that arose at the beginning of the century (§61). We consider some of them now.

2.1 Equational theories

Hilbert changed his basic attitude towards consistency proofs only around 1903 after the discovery of the elementary contradiction of Russell and Zermelo, which convinced him that there was a *deep* problem. In early 1904 he wrote to Adolf Hurwitz and claimed: 'exactly the most important and most interesting questions [concerning the foundations of arithmetic] have not been settled by Cantor and Dedekind (and a fortiori not by Weierstrass and Kronecker)'. He announced his intention to offer in the next semester a seminar on the 'logical foundations of mathematical thought' (original in [Dugac, 1976, 271]). The lecture notes from that term contain remarks on Dedekind's achievements, but insist that fundamental difficulties remain:

> He [Dedekind] arrived at the view that the standpoint of considering the integers as obvious cannot be sustained; he recognized that the difficulties Kronecker saw in the definition of irrationals arise already for integers; furthermore, if they are removed here, they disappear there. This work [*Was sind und was sollen die Zahlen?*] was epochal, but it did not yet provide something definitive, certain difficulties remain. These difficulties are connected, as for the definition of the irrationals, above all to the concept of the infinite; [. . .]

All of this set the stage for the talk of August 1904 at the International Congress of Mathematicians at Heidelberg. Hilbert [1905] stresses there the programmatic goal of developing logic and mathematics, in particular arithmetic, simultaneously. His theory of arithmetic is now restricted and deals only with natural numbers; it consists of axioms for identity and Dedekind's requirements for a *simply infinite system*, except that the induction principle is not formulated. The consistency of this purely equational system is established by an inductive argument on derivations. The work has real shortcomings, as there is neither a calculus for sentential logic nor a proper treatment of quantification. In sum, Hilbert initiates an important shift from *semantic* to *syntactic* arguments, but the formal set-up is inadequate as a framework for arithmetic, and the ultimate goal of the consistency proof remains to guarantee the existence of a set, here of the 'smallest infinite'.

Henri Poincaré (1854–1912) challenged the foundational import of Hilbert's considerations on account of the inductive character of the consistency proof [Poincaré, 1905–1906]. His incisive analysis shifted Hilbert's attention not away from foundational concerns (they are documented by lectures throughout the period from 1905 to 1917), but from the syntactic approach advocated in the Heidelberg talk. Indeed, under the impact of a detailed study of *Principia mathematica* beginning in 1913, Hilbert flirted again with logicism. What resulted from this study, very importantly as it contains the first exposition of modern mathematical logic, were the lectures 'Prinzipien der Mathematik', given in the winter semester of 1917–1918 with the assistance of Bernays. Their logicism was abandoned in the following year; a radical constructivism was adopted instead and subsequently aban-

doned; finally, the finitist consistency program was formulated in lectures given in the winter semester of 1921–1922.

2.2 Quantifier-free systems

The evolution towards this program started in the summer semester 1920, when Hilbert came back to the syntactic approach of his course of 1905. The notes from that semester contain a consistency proof for almost exactly the same fragment of arithmetic as that discussed in the Heidelberg talk; the modified argument is presented in the first part of [Hilbert, 1922] and its strategic point is made explicit there: 'Poincaré's objection, claiming that the principle of complete induction cannot be proved but by complete induction, has been refuted by my theory'.

In the second part of [Hilbert, 1922] the theory is expanded to include an appropriate logical calculus; he emphasizes that 'all formulas and statements of arithmetic can be obtained in a formal way'. The editors of his *Gesammelte Abhandlungen* mention that 'a schema for the introduction of functions by recursion equations' has to be added, if this last goal is to be reached. As to the claimed consistency result, they assert that it holds only if quantifiers are excluded and the induction axiom is replaced by the induction rule. With these modifications consistency is claimed though not proved there, for a theory that includes primitive recursive arithmetic. This work is the beginning of a genuinely new direction, which is best articulated in [Bernays, 1922] and given its principled formulation in Hilbert's Leipzig talk: the instrumental character of extensions that go beyond finitist mathematics is now emphasized.

The developments leading to a proof of the above result can be followed in contemporaneous lecture notes; the proof is only sketched in [Hilbert, 1923], but was given in detail during the winter semester of 1922–1923. The first step turns linear proofs into trees so that any formula occurrence is used at most once as a premise of an inference. That prepares the second step, namely, the elimination of all (necessarily free) variables through appropriate substitutions by a numeral. In the third step the numerical value of the closed terms and the truth-value of the formulas are determined. As all formulas in the final syntactic configuration turn out to be true, an inconsistency cannot be proved. Primitive recursively defined functions are admitted and treated in the argument. The rule of induction for quantifier-free formulas is also added, though not incorporated into the argument—it could be, as it was done already in 1921–1922.

From a contemporary perspective the arguments reveal something very important: as soon as a formal theory contains a class of finitist functions it is necessary to appeal to a wider class of functions in this kind of consistency proof. An *evaluation function* is needed to determine uniformly the numerical value of terms, and such a function is no longer in the given class. As the formal system considered in the above consistency proof includes primitive recursive arithmetic, the consistency proof goes beyond the means available in primitive recursive arithmetic. Finitist mathematics is consequently stronger than primitive recursive arithmetic at this early stage of proof theory. Indeed, as we will see, that assessment of the relative strength is clearly sustained throughout the development reported in this essay.

2.3 Quantifiers and ε-terms

The above proof theoretic considerations are preliminary in that they concern a theory that is *part* of finitist mathematics and thus *need* not be secured by a consistency proof. The truly expanding step involves theories with quantifiers treated according to Hilbert's *Ansatz*; that is indicated in [Hilbert, 1922] and elaborated in [Hilbert, 1923]. There, he sketches how quantifiers can be eliminated with the τ-function, the dual of the ε-operator, which replaces the τ-symbol in early 1923. The τ-function associates with every predicate $A(a)$ a particular object $\tau_x.A(x)$ or simply τA; it satisfies the *transfinite axiom* $A(\tau A) \to A(a)$ and allows the definition of the quantifiers:

$$(x)A(x) \leftrightarrow A(\tau A) \quad \text{and} \quad (Ex)A(x) \leftrightarrow A\big(\tau(\sim A)\big). \tag{1}$$

Hilbert extends the consistency argument to the 'first and simplest case' that goes beyond the finitist system and describes a particular process of eliminating instances of the transfinite axiom (later also called *epsilon axiom*, *epsilon formula* or *critical formula*).

The further development is quick and limited. Wilhelm Ackermann (1896–1962) directly continues Hilbert's proof-theoretic work in his thesis but modifies the elimination procedure for epsilon terms. His paper, based on the thesis, was submitted on 30 March 1924 and published in early 1925; it starts out in Section II with a concise review of Hilbert's considerations. That Section is entitled, tellingly, 'The consistency proof before the addition of the transfinite axioms'. At first it was believed that Ackermann [1925] had established the consistency of arithmetic and analysis; but a note was added 'in proof' restricting the result significantly. Von Neumann, whose paper 'Zur Hilbertschen Beweistheorie' was submitted on 29 July 1925, tried to clarify the extent of Ackermann's result and asserts that it covers Russell's mathematics without the axiom of reducibility or Hermann Weyl's system in his book *Das Kontinuum* (1918) [von Neumann, 1927, 46]. In his talk at the International Congress of Mathematicians held in Bologna in 1928, Hilbert [1929] stated, quite in line with von Neumann's observation, that the consistency of full number theory had been secured by the proofs of Ackermann and von Neumann; according to Bernays in his preface to the second volume, that belief was sustained until 1930. Indicating the depth of Dedekind's influence, Hilbert formulated as Problem I of his Bologna talk the consistency of the ε-axioms for function variables and commented later: 'The solution of problem I justifies also Dedekind's ingenious considerations in his essay *Was sind und was sollen die Zahlen?*'.

As we know now and as was recognized in 1931, Ackermann and von Neumann had established only the consistency of arithmetic with quantifier-free induction. In late 1933 Gödel attributed the most far-reaching partial result in the pursuit of Hilbert's program still to Herbrand, who in [Herbrand, 1931] had extended the Ackermann/von Neumann result by allowing a larger class of finitist functions that included, in particular, the non-primitive recursive Ackermann function. By then, Herbrand knew of Gödel's incompleteness theorems and agreed with von Neumann's related assertion: 'If there is a finitist consistency proof at all, then it can be formalized. Thus, Gödel's proof implies the impossibility of a consistency proof'. The historical development as sketched above is actually reflected in the structure of *Grundlagen der Mathematik*, whose systematic metamathematical content is to be described in the next two sections.

3 THE FIRST VOLUME

According to the preface of this volume, a presentation of proof theory had almost been completed, when the publication of papers by Herbrand and Gödel in 1931 produced a deeply changed situation for proof theory. This resulted in an extension of the scope of the work and its division into two volumes. The volumes were completed in early 1934 and early 1939; though both volumes use much material from joint work in the 1920s, the actual writing of the volumes was done by Bernays. The contents of the volumes are summarized in Table 1.

The eight chapters of Volume I can be divided roughly into three parts: Chapters 1 and 2 introduce the central foundational issues, Chapters 3 to 5 develop systematically the logical framework of first-order logic (with identity) and Chapters 6 to 8 investigate the consistency problem and other metamathematical questions for a variety of (sub-) systems

Table 1. Summary by Chapters of *Grundlagen der Mathematik*. Titles translated.

Chapter; pp.	**Chapter title**
I.1; 19	The consistency problem in axiomatics as a logical decision problem.
I.2; 23	Elementary number theory. Finitist inference and its limits.
I.3; 23	The formalization of logical inference I; the propositional calculus.
I.4; 79	The formalization of logical inference II; the predicate calculus.
I.5; 46	Inclusion of identity. Completeness of the one-place predicate calculus.
I.6; 78	The consistency of infinite domains of individuals. Beginnings of number theory.
I.7; 97	Recursive definitions.
I.8; 76	The concept “that, which” and its eliminability.
II.1; 48	The method of elimination of bound variables by means of Hilbert’s ε-symbol.
II.2; 82	Proof theoretic investigation of number theory by means of methods connected with the ε-symbol.
II.3; 75	Application of the ε-symbol for the investigation of the logical formalism.
II.4; 48	The method of the arithmetization of metamathematics applied to the predicate calculus.
II.5; 120	The reason for extending of the methodological frame for proof theory.
II.Supp. I;16	Overview of the predicate calculus and connected formalisms.
II.Supp. II; 29	A sharpening of the concept of calculable function and Church’s theorem on the decision problem.
II.Supp. III; 58	On certain parts of the propositional calculus and their deductive demarcation by means of schemata.
II.Supp. IV; 44	Formalisms for the deductive development of analysis.

of number theory. Volume I focuses on the development of proof theory without use of the ε-operator.

3.1 Existential axiomatics

Chapter 1 begins with a general discussion of axiomatics, at the center of which is a distinction between *contentual* and *formal axiomatic theories*. This distinction occurs under different formulations throughout Hilbert and Bernays's writings. Contentual axiomatic theories (examples of which include Euclid's geometry, Newton's mechanics and Clausius's thermodynamics) draw on experience for the introduction of their fundamental concepts and basic principles, which are understood contentually. By contrast, formal axiomatic theories such as Hilbert's axiomatization of geometry abstract away such intuitive content; they begin with the assumption of a fixed system of things (or several such systems), which is delimited from the outset and constitutes a 'domain of individuals for all predicates from which the statements of the theory are built up' (p. 2). The assumption of the existence of such a domain of individuals constitutes an 'idealizing assumption that joins the assumptions formulated in the axioms' (p. 3). Hilbert and Bernays elsewhere refer to this approach as *existential axiomatics*. While they clearly consider formal axiomatics to be a sharpening of contentual axiomatics, nonetheless they are quite explicit that these two types of axiomatics complement each other and are both necessary.

Through a general discussion of the consistency problem for formal axiomatic theories, they are led to conclude that the consistency of a formal axiomatic theory with a finite domain can be established by the exhibition of a model satisfying that system; however, one cannot proceed in this fashion for formal axiomatic theories with infinite domains. Consistency proofs for such theories present a special problem, because 'reference to non-mathematical objects cannot settle the question whether an infinite manifold exists; the question must be solved within mathematics itself' (p. 17). One must treat, they argue, the consistency problem for a formal axiomatic theory F with an infinite domain as a logical problem. This involves i) the formalization of principles of logical reasoning for F, and ii) a proof that from F one cannot derive (using these principles) both a formula and its negation. In short, one must treat the consistency problem from a proof theoretic perspective.

Such a proof need not be given individually for each F. Instead, one need only carry out such a proof for some axiom system F that 1) has a structure that is sufficiently *surveyable* to make a consistency proof for the system plausible, and 2) has a rich enough structure so that by assuming the existence of a system S of things and relations satisfying F, one can derive the satisfiability of axiom systems for the branches of physics and geometry. The satisfiability of an axiom system from those subjects is to be accomplished by representing its objects by individuals (or complexes of individuals) of S and its basic relationships by predicates constructed from those of S by logical operations. Hilbert and Bernays identify arithmetic (including number theory and analysis) as a candidate for such an F.

3.2 Finitist considerations

For such a consistency argument to be foundationally significant, it must avoid the idealizing existence assumptions made by formal axiomatic theories. But if a proof-theoretic

justification of arithmetic by elementary means should be possible, might it not be possible to give a direct development of arithmetic free from non-elementary assumptions (and thus not requiring any additional foundational justification)?

The answer to this question involves elementary presentations of parts of number theory and formal algebra; these presentations simultaneously serve to introduce the *finitist standpoint*. The finitist deliberations take here their *purest form*, i.e. the form of '*thought experiments* involving objects assumed to be *concretely given*' (p. 20). The word 'finitist' is intended to convey the idea that a consideration, a claim or definition respects that objects are to be representable in principle, and that processes are to be executable in principle (p. 32).

Having given finitist presentations of elementary number theory and formal algebra, Hilbert and Bernays remark that one cannot obtain a direct, elementary justification for all of mathematics, because already in number theory and analysis one uses non-finitist principles. While it is conceivable one could circumvent the use of such principles in number theory (where one only assumes the existence of the domain of integers), the case is different for analysis. There one assumes in addition the existence of real numbers, that is, infinite sets of integers, and applies the principle of the excluded middle also to these extended domains.

Thus one is led back to the strategy of proceeding in an indirect fashion, i.e., of using proof theory as a tool to secure the consistency of mathematics. As part of this strategy, Hilbert and Bernays adopt the methodological requirement that proof theory be finitist. This requirement ensures that the sought after consistency proof for arithmetic will avoid making idealizing existential assumptions which, after all, are in need of justification. This requirement that proof theory be finitist is relaxed only at the end of the second volume when 'extensions of the methodological framework of proof theory' are considered.

The first stage of this endeavor, the formulation of an appropriate logical formalism, occupies Chapters 3–5. The logical systems they develop are so close to contemporary ones that we do not discuss them in detail; they can actually be traced back to the lectures given in 1917–1918 and are presented already in [Hilbert and Ackermann, 1928]. The systematic development of logical formalisms is accompanied by their proof theoretic investigation. For instance, these chapters contain a number of normal form results as well as a proof of the completeness of the monadic predicate calculus with identity.

3.3 Consistency proofs

The second stage, in Chapters 6 and 7, involves the formulation and investigation of subsystems of number theory, which can be arranged into two groups. The first group of systems consists of weak fragments of arithmetic containing first-order quantification but few, if any, function symbols. These formalisms extend the predicate calculus with equality by mathematical axioms for 0, successor and $<$; some of them also involve quantifier-free induction. Hilbert and Bernays explore relations between them and establish independence, as well as consistency results. The main technique for giving consistency proofs is that discussed in section 1.2. However, since the formalisms contain quantifiers, an additional procedure is required here, namely a reduction procedure that assigns quantifier-free formulas, *reducts* acting as witnesses, to formulas containing quantifiers. The method underlying this procedure is due to Herbrand and to Emil Presburger. Additionally, the procedure for the replacement of free variables now must also handle free formula variables.

A further difference is that the consistency results are inferred from more general results involving the notion of *verifiability*, which is an extension of the notion of truth to certain formulas containing free variables, bound variables, and recursively defined function signs. More precisely, letting A be a formula of the formalism F,

i) if A is a numeric formula (that is, if it is composed of equalities and inequalities between numerals by means of sentential connectives), then it is verifiable if it is true;

ii) if A contains free numeric variables (but no formula variables or bound variables), then it is verifiable if one can show by finitist means that the substitution of arbitrary numerals for variables (followed by the evaluation of all function-expressions and their replacement through their numerical values) yields a true numeric formula;

iii) if A contains bound variables but no formula variables, then it is verifiable if its reduct is verifiable (according to i) and ii)).

In order to establish the consistency of a formalism F, one proves now that every formula not containing formula variables is verifiable, if it is derivable in F. Since $0 \neq 0$ is not verifiable, it is not derivable in F; it follows that F is consistent.

The second group of subsystems of number theory contains formalisms arising from the elementary calculus with free variables (the quantifier-free fragment of the predicate calculus) through the addition of functions defined by primitive recursion. Hilbert and Bernays start Chapter 7 with a discussion of the formalization of the principle of definition by recursion. They take the simplest schema of recursion to be

$$f(a, \ldots, k, 0) = \mathrm{a}(a, \ldots, k), \tag{2}$$

$$f(a, \ldots, k, n') = \mathrm{b}\big(a, \ldots, k, n, f(a, \ldots, k, n)\big), \tag{3}$$

where a and b denote previously defined functions and $a, \ldots, k, n$ are numerical variables. After discussing this definitional principle, they prove a

GENERAL CONSISTENCY THEOREM. *Let F be a formalism extending the elementary calculus with free variables by verifiable axioms* (*that may contain recursively defined functions whose defining equations are taken as axioms*) *and the schema of quantifier free induction, then every derivable formula of F is verifiable.*

They explicitly take this theorem to establish the consistency of a number of formalisms including that of recursive number theory, which they develop at length in order to illustrate the strength of recursive definitions. As their notion of recursive number theory is equivalent to primitive recursive arithmetic, finitist mathematics here goes beyond primitive recursive arithmetic. Following this development they discuss formalisms arising from the extension of the recursion and induction schemas and remark that their previous consistency results are easily extended to these systems as well; these remarks push the bounds of finitist mathematics still further.

3.4 Full number theory

The third stage of the development carried out in the first volume occurs towards the end of Chapter 7 and in Chapter 8. Here one finds a third group of formalisms that are each equivalent to full Peano Arithmetic. The first of these is the formalism of the axiom system (Z); call this formalism Z. When arriving at Z, Hilbert and Bernays comment that the techniques used in their previous consistency proofs for fragments of number theory cannot be generalized to Z. The problem is that any reduction procedure for Z would provide a decision procedure for Z and thus would allow one to solve all problems of number theory. They leave the possibility of such a procedure as an open problem (whose solution, if it exists, is a long way off) and focus on showing that Z provides the means for the formalization of full number theory.

With this end in mind, Hilbert and Bernays prove in Chapter 8 that all recursive functions are representable in Z. This proof involves establishing three separate claims: 1) that the least number operator μ can be explicitly defined in terms of Russell and Whitehead's ι-symbol; 2) that any recursive definition (a notion that they leave unanalyzed) can be explicitly defined in Z_μ (that is, Z extended by defining axioms for the μ-operator); and 3) that the addition of the ι-rule to Z is a conservative extension of Z. After the discussion of some additional results, such as the general eliminability of function symbols using predicate symbols, the first volume concludes with the remark that the above results entail the consistency of Z_μ relative to that of Z, but that none of the results or methods considered so far suffice to show that Z is consistent.

4 THE SECOND VOLUME

The second volume picks up where the first left off. It presents in Chapters 1 and 2 Hilbert's proof theoretic 'Ansätze' based on the ε-symbol as well as related consistency proofs; this is the first main topic. The methods used there open a simple approach to Herbrand's theorem, which is at the center of Chapter 3. The discussion of the decision problem at the end of that chapter leads, after a thorough discussion of the 'method of the arithmetization of metamathematics', in the next chapter to a proof theoretic sharpening of Gödel's completeness theorem. The remainder of the volume is devoted to the second main topic, the examination of the fact, which is the basis for the necessity to expand the frame of the contentual inference methods, which are admitted for proof theory, beyond the earlier delimitation of the 'finitist standpoint'. Of course, Gödel's incompleteness theorems are at the center of that discussion.

4.1 Limited results

The consistency proofs in Section 7.a) of the first volume were given for quantifier-free systems. Now these theories are embedded in the system of full predicate logic together with the ε-axioms, which have the form $A(a) \to A(\varepsilon_x.A(x))$; the ε-terms $\varepsilon_x.A(x)$ represent individuals having the property expressed by $A(a)$, if the latter holds of any individual at all. The crucial task is to eliminate all references to bound variables from proofs of theorems that do not contain them; axioms used in these proofs must not contain bound

variables either. In the formulation of Hilbert and Bernays, the consistency of a system of proper axioms relative to the predicate calculus together with the ε-axioms is to be reduced to the consistency of the system relative to the elementary calculus (with free variables) (p. 33). The consistency of the latter system is recognized on account of a suitable finitist interpretation. Thus, Hilbert and Bernays emphasize that operating with the ε-symbol can be viewed as 'merely an auxiliary calculus, which is of considerable advantage for many metamathematical considerations' (pp. 12–13).

In the framework of the extended calculus, bound variables can be seen to be associated really only with terms, as the quantifiers can be defined in a way dual to that shown earlier for the τ-symbol. The initial elimination result is the

FIRST ε-THEOREM. *If the axioms $A_1, \ldots, A_k$ and the conclusion of a proof do not contain bound individual variables or* (*free*) *formula variables, then all bound variables can be eliminated from the proof.*

The argument can be extended to cover proofs of purely existential formulas, but the formal proofs then yield as their conclusion a suitable disjunction of instances of the existential formula. Based on this extension Hilbert and Bernays prove their

CONSISTENCY THEOREM. *If the axioms $A_1, \ldots, A_k$ are verifiable, then* i) *any provable formula containing at most free individual variables is verifiable, and* ii) *for any provable, purely existential formula $(Ex_1)\cdots(Ex_n)\ A(x_1, \ldots, x_n)$* (*with only the variables shown*) *there are variable-free terms $t_1, \ldots, t_n$ such that $A(t_1, \ldots, t_n)$ is true.*

This theorem is applied to establish the consistency i) of Euclidean and Non-Euclidean geometry without continuity assumptions in section 1.4, and ii) of arithmetic with recursive definitions, but only quantifier-free induction as in sections 2.1 and 2.2. In essence then, the consistency theorem from [Herbrand, 1931] has been reestablished in a subtly more general way, as is emphasized on p. 52: Hilbert and Bernays allow the introduction of a larger class of recursive functions. We can put the result also in a different historical context and see that the consistency proof of 1923 for the quantifier-free system of primitive recursive arithmetic has been extended to cover that system's expansion by full classical quantification theory.

The remainder of Chapter 2 discusses the difficulty of extending the elimination procedure (in the proof of the first ε-theorem) to a system with full induction and examines Hilbert's original *Ansatz* for eliminating ε-symbols. (As to the character of the original and the later version of the elimination method and Ackermann's work see pp. 21, 29, 92ff, the note on p. 121, as well as Bernays's preface.) The next two chapters investigate the formalism for predicate logic, beginning in Chapter 3 with a proof of the

SECOND ε-THEOREM. *If the axioms and the conclusion of a proof* (*in predicate logic with identity*) *do not contain ε-symbols, then all ε-symbols can be eliminated from the proof.*

Then Herbrand's theorem is obtained as well as a variety of criteria for the refutability of formulas in predicate logic; proofs of the Löwenheim–Skolem theorem and of Gödel's completeness theorem are also given. These considerations are used to establish results concerning the decision problem, and solvable cases as well as reduction classes are discussed. In Chapter 4 Gödel's method of the 'arithmetization of metamathematics' is presented in great detail and applied to obtain a fully formalized proof of the completeness theorem.

Here is one standard formulation of the completeness theorem: consistency of an axiom system relative to the calculus of predicate logic coincides with satisfiability of the system by an arithmetic model. The formalized proof is intended to establish a kind of finitist equivalent (p. 205) to a consequence of this formulation, namely, that the consistency relative to the predicate calculus guarantees consistency in an open contentual sense ('im unbegrenzten inhaltlichen Sinne'). The finitist equivalent is formulated in terms of irrefutability roughly as follows: if a formula is irrefutable in predicate logic, then it remains irrefutable in 'every consistent number theoretic formalism', that is, in every formalism that is consistent and remains consistent when the axioms of Z_μ and possibly also verifiable formulas are added (p. 253). That fact can be interpreted as expressing a deductive closure of the predicate calculus, but obviously only if Z_μ is consistent. Thus, there is an additional reason for establishing the consistency of this number theoretic formalism.

4.2 Incompleteness

The discussion of Gödel's incompleteness theorems (§71) begins with a thorough investigation of semantic paradoxes. However, this investigation does not try to 'solve' the paradoxes in the case of natural languages, but focuses on the question under what conditions analogous situations can occur in the case of *formalized languages*. These conditions are formulated quasi-axiomatically for general deductive formalisms F taking for granted that there is a bijection between the expressions of F and natural numbers, a 'Gödel-numbering'. The formalism F and the numbering are required to satisfy roughly two *representability conditions*: R1) primitive recursive arithmetic is 'contained in' F; and R2) the syntactic properties and relations of F's expressions, as well as the processes that can be carried out on such expressions, are given by primitive recursive predicates and functions.

For the consideration of the first incompleteness theorem the second representability condition is made more specific. It now requires that the *substitution function* (yielding the number of the expression obtained from an expression with number **k**, when every occurrence of the number variable a is replaced by a numeral **l**) is given primitive recursively by a binary function $s(k,l)$ and the *proof predicate* by a binary relation $B(m,n)$ (holding when **m** is the number of a sequence of formulas constituting an F-derivation of the formula with number **n**). Consider, as Gödel did, the formula $\sim B(m, s(a,a))$; according to the first representability condition this is a formula of the formalism F and has a number, say **p**. Because of the defining property of $s(k,l)$, the value of $s(\mathbf{p},\mathbf{p})$ is then the number **q** of the formula $\sim B(m, s(\mathbf{p},\mathbf{p}))$. The equation $s(\mathbf{p},\mathbf{p}) = \mathbf{q}$ is provable in F; thus, $\sim B(m, s(\mathbf{p},\mathbf{p}))$ is actually equivalent to $\sim B(m,\mathbf{q})$ and expresses that 'the formula with number **q** is not provable in F'. As **q** is the number of $\sim B(m, s(\mathbf{p},\mathbf{p}))$, this formula consequently expresses (via the equivalence) its own underivability. The argument adapted from

that for the liar paradox leads, from the assumption that this formula is provable, directly to a contradiction in F. But instead of encountering a paradox, we infer now that the formula is not provable, if the formalism F is consistent.

Hilbert and Bernays discuss—following Gödel and assuming the ω-consistency of F—the unprovability of the sentence $\sim(x)\sim B(m, \mathbf{q})$. Then they establish the Rosser version of the first incompleteness theorem, i.e., the independence of a formula R from F assuming just F's consistency. Thus, a 'sharpened version' of the theorem can be formulated for deductive formalisms satisfying certain conditions: 'One can always determine a unary primitive recursive function f, such the equation $f(m) = 0$ is not provable in F, while for each numeral $\mathbf{l}$ the equation $f(\mathbf{l}) = 0$ is true and provable in F; neither the formula $(x)f(x) = 0$ nor its negation is provable in F' (p. 279). This sharpened version of the theorem asserts that every sufficiently expressive, sharply delimited, and consistent formalism is deductively incomplete. An important consequence of this result is discussed in section 5.1.

4.3 Unprovability of consistency

For a formalism F that is consistent and satisfies the restrictive conditions, the proof of the first incompleteness theorem shows the formula $\sim B(m, \mathbf{q})$ to be unprovable. However, it also shows that the sentence $\sim B(\mathbf{m}, \mathbf{q})$ holds and is provable in F, for each numeral $\mathbf{m}$. The second incompleteness theorem is obtained by formalizing these considerations, i.e. by proving in F the formula $\sim B(m, \mathbf{q})$ from the formal expression C of F's consistency. That is possible, however, only if F satisfies certain additional conditions, the so-called *derivability conditions*. Hilbert and Bernays conclude immediately 'in case the formalism F is consistent no formalized proof of this consistency, i.e. no derivation of that formula C, can exist in F' (p. 284).

The formalized argument makes use of the representability conditions R1) and R2), where the second condition now requires also that there is a unary primitive recursive function e, which when applied to the number $\mathbf{n}$ of a formula yields as its value the number of the negation of the formula. These then are the derivability conditions: D1) If there is a derivation of a formula with number $\mathbf{l}$ from a formula with number $\mathbf{k}$, then the formula $(Ex)B(x, \mathbf{k}) \to (Ex)B(x, \mathbf{l})$ is provable in F; D2) The formula $(Ex)B(x, e(k)) \to (Ex)B(x, e(s(k, l)))$ is provable in F; and D3) If $f(m)$ is a primitive recursive term with m as its only variable and if $\mathbf{r}$ is the number of the equation $f(a) = 0$, then the formula $f(m) = 0 \to (Ex)B(x, s(\mathbf{r}, m))$ is provable in F. Consistency is formally expressed by $(Ex)B(x, n) \to \sim(Ex)B(x, e(n))$; starting with that assumption, the formula $\sim B(m, \mathbf{q})$ is obtained in F by a rather direct argument on pp. 286–288.

There are two brief remarks with which we want to complement this metamathematical discussion of the incompleteness theorems. The first simply states that verifying the representability conditions and the derivability conditions is the central mathematical work that has to be done; Hilbert and Bernays accomplish this for the formalism Z_μ (starting on p. 293) and for Z (beginning on p. 324). Thus, the second volume of *Grundlagen der Mathematik* contains the first full argument for the second incompleteness theorem; after all, Gödel's paper contains only a minimal sketch of a proof. However, it has to added—and

that is the second brief remark—that the considerations are not fully satisfactory for a *general* formulation of the theorems, as there is no argument given why deductive formalisms should satisfy the particular restrictive conditions on their syntax. This added observation points to one of the general methodological issues discussed next.

5 PHILOSOPHICAL AND MATHEMATICAL ISSUES

The existential formal axiomatics that emerged in the second half of the 19th century and found its remarkable expression in Hilbert's *Grundlagen der Geometrie* (§55) constituted the major pressing issue for the various Hilbert programs during the period from 1899 to 1934, the date of the publication of the first volume of *Grundlagen der Mathematik*. The finitist consistency program that began to be pursued in 1922 is the intellectual thread holding the investigations in both volumes together. The general programmatic direction was formulated clearly in the first volume and presented above in section 2.1. The ultimate goal of proof theoretic investigations, as Hilbert formulated it in the preface to volume I, is to recognize the usual methods of mathematics, without exception, as consistent. Hilbert continued, 'With respect to this goal I would like to emphasize the following: the view, which temporarily arose and maintained that certain recent results of Gödel imply the infeasibility of my program, has been shown to be erroneous'. How is the program affected by those results? Is it indeed the case, as Hilbert expressed it also in 1934, that the Gödel theorems just force proof theorists to exploit the finitist standpoint in a sharper way?

5.1 Issue of completeness

The second question is raised prima facie only through the second incompleteness theorem. However, Hilbert and Bernays discuss also the effect of the first incompleteness theorem and ask quite explicitly (p. 280), whether the deductive completeness of formalisms is a necessary feature for the consistency program to make sense. They touched on this very issue already in pre-Gödel publications, Hilbert in his Bologna Lecture of 1928 and Bernays in his penetrating article [1930]. He formulated in his lecture the question of the syntactic completeness for number theory and analysis as Problem III, and concluded the discussion by suggesting that 'in höheren Gebieten' (higher than number theory) it is thinkable that a system of axioms could be consistently extended by a statement S, but also by its negation $\sim S$; the acceptance of one of the statements is then to be justified by 'systematic advantages (principle of the permanence of laws, possibilities of further developments etc.)'.

Hilbert conjectured that number theory is deductively complete (p. 59). That is reiterated in [Bernays, 1930] and followed by the remark that 'the problem of a real proof for this is completely unresolved'. The problem becomes even more difficult, Bernays continues, when we consider systems for analysis or set theory. However, this 'Problematik' is not to be taken as an objection against the standpoint presented (p. 59):

> We only have to realize that the [syntactic] formalism of statements and proofs we use to represent our conceptions does not coincide with the [mathematical] formalism of the structure we intend in our thinking. The [syntactic] formalism

> suffices to formulate our ideas of infinite manifolds and to draw the logical consequences from them, but in general it [the syntactic formalism] cannot combinatorially generate the manifold as it were out of itself.

That is also the central point in the general discussion of the first incompleteness theorem (p. 280). Indeed, Hilbert and Bernays emphasize there that in formulating the problems and goals of proof theory they avoided from the very beginning 'to introduce the idea of a total system for mathematics with a philosophically principled significance'. It suffices for their purposes to characterize the actual systematic structure of analysis and set theory in such a way that it provides an appropriate frame for (the reducibility of) the geometric and physical disciplines.

From these reflective remarks it follows that the first incompleteness theorem for the central formalisms F (of number theory, analysis, and set theory) does not directly undermine Hilbert's program. Nevertheless, it raises in its sharpened form a peculiar issue: any finitist consistency proof for F would yield a finitist proof of a statement in recursive number theory—that is not provable in F. Finitist methods would thus go beyond those of analysis and set theory, even for the proof of number theoretic statements. This is a 'paradoxical' situation, in particular, as Hilbert and Bernays quite unambiguously state in the first volume (p. 42), 'finitist methods are included in the usual arithmetic'. Consequently, even the first theorem forces us to address two general tasks, namely, i) to explore the extent of finitist methods, and ii) to demarcate appropriately the methodological standpoint for proof theory.

5.2 The extent of finitist methods

Tasks i) and ii) are usually associated with the second incompleteness theorem, which, as emphasized at the end of Section 4.3, allows us to infer directly and sharply that a finitist consistency proof for a formalism F (satisfying the representability and derivability conditions) cannot be carried out in F. Hilbert and Bernays explore the extent of finitist methods in Section 5.3.a) by first trying to answer the question, in which formalism their various finitist investigations can actually be carried out. The immediate claim is that most considerations can be formalized, perhaps with a great deal of effort, in primitive recursive arithmetic (p. 340). But then they assert: 'At various places this formalism is admittedly no longer sufficient for the desired formalization. However, in each of these cases the formalization is possible in Z_μ'. They point to the more general recursion principles from Chapter 7 of the first volume as an example of 'procedures of finitist mathematics' that cannot be captured in primitive recursive arithmetic, but can be formalized in Z_μ.

In the remainder of Section 5.3.a) they discuss 'certain other typical cases', in which the boundaries of primitive recursive arithmetic are too narrow to allow a formalization of their prior finitist investigations. There is, first of all, the issue of an evaluation function that is needed for the consistency proof of primitive recursive arithmetic (already in volume I) but cannot be defined by primitive recursion (p. 341). Secondly, there is the general concept of a calculable function (p. 342); that concept is used (p. 189) to formulate a finitistically sharpened notion of satisfiability, i.e. *effective satisfiability*, in finitist treatments of solvable cases of the decision problem. Thirdly, they discuss the principle of induction

for universally quantified formulas used in consistency proofs (p. 344). The issue surrounding this principle is settled metamathematically, as we now know, by later proof theoretic work: the system of elementary number theory with this induction principle is conservative over primitive recursive arithmetic.

As to ii), some remarks concerning supplement II are relevant in the above context, as the notion of a calculable function has to be sharpened in such a way that it can be formalized. The presentation in that supplement of the negative solution of the decision problem is preceded by a conceptual analysis of the concept 'reckonable function', i.e. of a function whose values can be calculated according to rules. The latter rather vague notion is sharpened, in a way that is methodologically very similar to the analysis of the incompleteness theorems, namely by formulating *recursiveness conditions* for deductive formalisms that allow equational reasoning. The central condition requires the proof predicate to be primitive recursive. It is then shown that the functions calculable in formalisms satisfying the recursiveness conditions are exactly the general recursive ones. The latter notion can be defined in the language of number theory as is necessary for the formalization in ii). Though the conceptual analysis is not fully satisfactory for the reason mentioned in Section 4.3, it is nevertheless a major and concluding step in the analysis of effectively calculable functions as pursued in the mid-1930s by Gödel, Alonzo Church, Stephen Kleene, and others.

5.3 Beyond finitism?

The examination of their own proof-theoretic practice leads Hilbert and Bernays to the conclusion that some considerations require means that go beyond primitive recursive arithmetic, but can be formally captured in Z_μ. It is at exactly this point that the second incompleteness theorem provides, as the title of Chapter 5 states, the 'reason for extending the methodological frame for proof theory'. Already on p. 253, as a transition from Chapter 4 to Chapter 5, Hilbert and Bernays state specifically that consequences of the theorem force us to view the domain of the contentual inference methods used for the investigations of proof theory more broadly 'than it corresponds to our development of the finitist standpoint so far'.

The question is, whether there are any methods that can still be called properly 'finitist' and yet go beyond Z_μ. Hilbert and Bernays argue that this is not a precise question, as 'finitist' is not a sharply delimited notion, but rather indicates methodological guidelines that enable us to recognize some considerations as definitely finitist and others as definitely non-finitist. The limits of finitist considerations are to be 'loosened' (vol. 2, 348), and two possibilities of such loosenings are considered that are quickly seen to be 'conservative'. Which further loosenings are 'admissible, if we want to adhere to the fundamental tendencies of proof theory?' Against this background two then recent results are examined: the reduction of classical arithmetic Z to the system $\mathscr{Z}$ of arithmetic with just minimal logic, and Gentzen's consistency proof for a version of $\mathscr{Z}$ (and thus of Z) using a special form of transfinite induction.

The reductive result that Hilbert and Bernays formulate is a slightly stronger one than that obtained by Gödel and, independently, by Gentzen. The proof showing that Z is consistent relative to $\mathscr{Z}$ is an elementary finitist one. Thus, the obstacle for obtaining a finitist consistency proof for Z does not lie in the fact that it contains the typically non-finitist

logical principles like *tertium non datur*! The obstacle appears already when one tries to give a finitist consistency proof for $\mathscr{Z}$. The consistency of Z would be established on the basis of any assumptions, 'which suffice to give a verifying interpretation of the restricted formalism' (p. 357). Such a contentual verification, based on interpretations of A.N. Kolmogoroff and Arend Heyting, is then examined with the conclusion that it involves the intuitionistic understanding of negation as absurdity.

The question is raised, whether—in a proof of the consistency of Z—the systematic use of absurdity could be avoided, as well as the appeal to an interpretation of the formalism (in contrast to its direct proof theoretic examination). It is claimed that Gentzen's consistency proof addresses both these issues. After a thorough discussion of the details of the system of ordinal notation and the (justification of the) principle of transfinite induction, but only the briefest indication of the structure of Gentzen's proof, the main body of the book concludes with some extremely general remarks about the significance of Gentzen's proof: it provides a perspective for the proof theoretic investigation also of stronger formalisms, when one clearly has to countenance the use of larger and larger ordinals. The volume concludes with the sentence: 'If this perspective should prove its value, then Gentzen's consistency proof would open a new phase of proof theory'. In this way, it seems, Bernays sees Gentzen's approach as overcoming 'the temporary fiasco of proof theory' he discussed in the introduction to volume II and attributed to 'exaggerated methodological demands put on the theory'.

No explicit final and definitive judgment on the (non-)finitist character of these two consistency proofs is actually articulated in the book. However, in the first volume (p. 43), intuitionism is viewed as a proper extension of finitist mathematics. That view is also expressed in contemporaneous papers by Bernays and in many later comments, perhaps most dramatically in his article on Hilbert, where the above relative consistency proof for Z is seen as the reason for the recognition 'that intuitionistic reasoning is not identical with finitist reasoning, contrary to the prevailing views at the time' [Bernays, 1967, 502]. As to Gentzen's consistency proof, Bernays states in the introduction to the second edition of volume II that the transfinite induction principle used in it is 'a non-finitist tool'.

5.4 Demarcation

In the introduction of the first edition and the detailed discussion there is perhaps an ambiguity; whether the extension of the finitist standpoint necessitated by the incompleteness theorems still is essentially the finitist standpoint as articulated in the first two chapters of volume I, or whether it is a proper extension compatible with the broader strategic considerations underlying proof theory. We think the ambiguity should be resolved in the latter sense; after all, the considerations in Chapter 5 come under the heading 'Transcending the former methodological standpoint of proof theory—Consistency proofs for the full number theoretic formalism'.

However, there is not even a broad demarcation of a new, wider methodological standpoint for proof theory; a reason for this lack is perhaps implicit in the remarks connecting the consistency proof for Z relative to intuitionistic arithmetic with Gentzen's consistency

proof (p. 360). It is claimed, first of all, that it is 'unsatisfactory from the standpoint of proof theory' to have only a consistency proof for Z that 'rests mainly on an interpretation of a formalism'. It is observed, secondly, that the only method of going beyond the formalism Z has been the formulation of truth definitions: a classical truth definition is given for Z on pp. 329–340, and the formalization of the consistency proof based on an intuitionistic interpretation would amount to using a truth definition. Thirdly and finally, it is argued that a consistency proof is desirable that rests on 'the direct treatment of the formalism itself'; that is seen in analogy for obtaining the consistency of primitive recursive arithmetic, where Hilbert and Bernays were not satisfied with the possibility of a finitist interpretation, but rather convinced themselves of the consistency by specific proof theoretic methods. Where in this discussion is even an opening for a broader demarcation?

6 CONCLUDING REMARKS

The free and open way in which Hilbert and Bernays joined in the 1920s a number of different tendencies into a sharply focused program with a special mathematical and philosophical perspective is remarkable. The program has been transformed, in accord with the broad strategy underlying Hilbert's proposal, to a *general reductive* one; here one tries to give consistency proofs for strong classical theories relative to 'appropriate constructive' ones. The expanding development of proof theory is one effect of Hilbert's broad view on foundational problems and of his sharply articulated questions. Another effect is visible in the rich and varied results of Hilbert, Bernays, and other members of the Hilbert School (Ackermann, Gentzen, Kurt Schütte); finally, we have to consider the stimulus his approach and questions provided to contemporaries outside the school (von Neumann, Herbrand, Gödel, Church, and Alan Turing). Indeed, there is no foundational enterprise with a more profound and far-reaching effect on the emergence and development of mathematical logic. What Ackermann [1934] formulated in his review of just the first volume, holds even more for the complete two-volume work, namely, that it 'is to be viewed in a line with the great publications of Frege, Peano, and Russell–Whitehead'.

NOTE

Ravaglia wrote section 3, Sieg the rest.

BIBLIOGRAPHY

The notes of the lectures given by Hilbert (and Bernays) are all available in the Hilbert *Nachlaß* at the University of Göttingen. They will be published in volume 2 and 3 of 'David Hilbert's lectures on the foundations of mathematics and physics, 1891 to 1933'.
Reviews of the book are gathered together at the end.

Ackermann, W. 1925. 'Begründung des "tertium non datur" mittels der Hilbertschen Theorie der Widerspruchsfreiheit', *Mathematische Annalen*, *93*, 1–36.

Bernays, P. 1922. 'Über Hilberts Gedanken zur Grundlegung der Mathematik', *Jahresbericht der Deutschen Mathematiker-Vereinigung*, *31*, 10–19.

Bernays, P. 1930. 'Die Philosophie der Mathematik und die Hilbertsche Beweistheorie', in *Abhandlungen zur Philosophie der Mathematik*, Darmstadt: Wissenschaftliche Buchgesellschaft, 1976, 17–61.

Dedekind, R. 1888. *Was sind und was sollen die Zahlen?*, in *Gesammelte mathematische Werke*, vol. 3, Braunschweig: Vieweg, 1932, 335–391. [See § 47.]

Dugac, P. 1976. *Richard Dedekind et les fondements des mathématiques*, Paris: Vrin.

Ewald, W. (ed.) 1996. *From Kant to Hilbert. A source book in the foundations of mathematics*, 2 vols., New York: Oxford University Press.

Gödel, K. 1931. 'Über formal unentscheidbare Sätze der *Principia Mathematica* und verwandter Systems I', *Monatshefte für Mathematik und Physik*, *38*, 173–198. [See §71.]

Gödel, K. 1933. 'The present situation in the foundations of mathematics', in *Collected works*, vol. 3, New York: Oxford University Press, 1995, 36–53.

Herbrand, J. 1931 'Sur la non-contradiction de l'arithmétique', *Journal für die reine und angewandte Mathematik*, *166*, 1–8.

Hilbert, D. 1899. *Grundlagen der Geometrie*, 1st ed., Leipzig: Teubner. [See §55.]

Hilbert, D. 1900. 'Über den Zahlbegriff', *Jahresbericht der Deutschen Mathematiker-Vereinigung*, *8*, 180–194.

Hilbert, D. 1901. 'Mathematische Probleme. Vortrag, gehalten auf dem internationalen Mathematiker-Kongress zu Paris 1900', *Archiv der Mathematik und Physik*, *(3) 1*, 44–63, 213–237. [See § 57.]

Hilbert, D. 1905. 'Über die Grundlagen der Logik und der Arithmetik', in *Verhandlungen des Dritten Internationalen Mathematiker-Kongresses*, Leipzig: Teubner, 174–185. [English trans. in [van Heijenoort, 1967], 129–138.]

Hilbert, D. 1922. 'Neubegründung der Mathematik', *Abhandlungen aus dem mathematischen Seminar der Hamburgischen Universität*, *1*, 157–177.

Hilbert, D. 1923. 'Die logischen Grundlagen der Mathematik', *Mathematische Annalen*, *88*, 151–165.

Hilbert, D. 1929. 'Probleme der Grundlegung der Mathematik', *Mathematische Annalen*, *102*, 1–9.

Hilbert, D. and Ackermann, W. 1928. *Grundzüge der theoretischen Logik*, 1st ed., Berlin: Springer.

Poincaré, H. 1905–1906. 'Les mathématiques et la logique', *Revue de métaphysique et de morale*, *13*, 815–835; *14*, 17–34, 294–317. [English trans. in [Ewald, 1996], vol. 2, 1021–1071.]

Toepell, M.-M. 1986. *Über die Entstehung von David Hilberts, Grundlagen der Geometrie*, Göttingen: Vandenhoeck & Ruprecht.

van Heijenoort, J. (ed.) 1967. *From Frege to Gödel. A source book in mathematical logic, 1879–1931*, Cambridge, MA: Harvard University Press.

von Neumann, J. 1927. 'Zur Hilbertschen Beweistheorie', *Mathematische Zeitschrift*, *26*, 1–46.

Reviews of the book consulted

Ackermann, W. 1934. Of volume 1, *Jahrbuch über die Fortschritte der Mathematik*, *60*, 17–19.

Black, M. 1940. Of volume 2, *Mind*, *49*, 239–248.

Carnap, R. 1939. Of volume 1, *The journal of unified science* (*Erkenntnis*), *8*, 184–187.

Kleene, S.C. 1940. Of volume 2, *Journal of symbolic logic*, *5*, 16–20.

Kalmar, L. 1938, 1941. Of both volumes, *Acta scientiarum mathematicarum*, *7*, 255; *10*, 79–80.

Kneebone, G.T. 1970. Of volume 1, second edition, *Journal of symbolic logic*, *35*, 321–323.

LIST OF AUTHORS

Titles, short titles or descriptions of the writings are appended to the affiliations.

FRANCK ACHARD University of Paris 8, France: 1873 Maxwell, *Electricity and magnetism*

DAVID AUBIN Institute of Mathematics, University of Paris 6, France: 1927 Birkhoff, *Dynamical systems*

JUNE BARROW-GREEN Faculty of Mathematics, Open University, Milton Keynes, Britain: 1890 Poincaré, memoir on three-body problem

DENIS BAYART Ecole Polytechnique, Paris, France: 1931 Shewhart, *Economic quality control*

LAURIE BROWN Department of Physics and Astronomy, Northwestern University, Evanston, Illinois, USA: 1930 Dirac, *Quantum mechanics* and 1932 von Neumann, *Quantenmechanik*

JOAO CARAMALHO DOMINGUES Department of Mathematics, University of Minho, Braga, Portugal: 1797–1800 Lacroix, *Traité du calcul*

ALBERTO CONTE Department of Mathematics, University of Turin, Italy: 1915–1934 Enriques and Chisini, book on algebraic functions

ROGER COOKE Department of Mathematics, University of Vermont, Burlington, USA; 1826 Abel, paper on the quintic equation; 1829 Jacobi, *Functionum ellipticarum*; 1872 Dedekind, *Stetigkeit und irrationalen Zahlen*; 1904–1906 Lebesgue, *Intégration* and *Séries trigonométriques*, and Baire, *Fonctions*; and 1932 Bochner, *Vorlesungen über Fouriersche Integralen*

LEO CORRY Cohn Institute for History and Philosophy of Science, Tel-Aviv University, Israel: 1895–1896 Weber, *Lehrbuch der Algebra*

ALAIN COSTE Girard Desargues Institute, University of Lyon 1, France: 1799–1802 Montucla, *Histoire des mathématiques*

PIERRE CRÉPEL 'Maply' Laboratory, CNRS and University of Lyon 1, France: 1743 d'Alembert, *Dynamique*; 1799–1802 Montucla, *Histoire des mathématiques*

TONY CRILLY Middlesex University Business School, London, Britain: 1923–1926 Brouwer and Urysohn, papers on dimension theory

ANDREW DALE School of Mathematical and Statistical Sciences, University of Natal, Durban, South Africa: 1764 Bayes, paper on probability theory

JOSEPH DAUBEN Graduate Center, City University of New York, USA: 1883 Georg Cantor, paper on set theory

SERGEI DEMIDOV Institute of the History of Science and Technology, University of Moscow, Russia: 1755 Euler, *Differentialis*

A.W.F. EDWARDS Gonville and Caius College, Cambridge, Britain: 1925 Fisher, *Statistical methods*

DELLA FENSTER Department of History, University of Richmond, Virginia, USA: 1919–1923 Dickson, *Number theory*

JOSÉ FERREIRÓS Department of Philosophy and Logic, University of Seville, Spain: 1888 Dedekind, *Was sind . . . Zahlen?* and 1889 Peano, *Arithmetices*

CRAIG FRASER Institute for the History and Philosophy of Science and Technology, University of Toronto, Canada: 1744 Euler, *Methodus inveniendi*; and 1797 Lagrange, *Fonctions analytiques*

I. GRATTAN-GUINNESS Middlesex University Business School, London, Britain: Introduction; 1796–1827 Laplace, *Exposition* and *Mécanique céleste*; Cauchy 1821, *Cours d'analyse* and 1823, *Résumé* on the calculus; 1822 Fourier, *Théorie analytique de la chaleur*; 1828 Green, *Electricity and magnetism*; 1854 Boole, *Laws of thought*; and 1910–1913 Whitehead and Russell, *Principia mathematica*

CATHERINE GOLDSTEIN CNRS—Institut de mathématiques de Jussieu, Paris, France: 1863 Dirichlet, *Vorlesungen über Zahlentheorie*

JEREMY GRAY Faculty of Mathematics, Open University, Milton Keynes, Britain: 1822 Poncelet, projective geometry; 1867 Riemann, thesis on geometries; and 1872 Klein, Erlangen programme

NICCOLO' GUICCIARDINI Department of Philosophy, University of Siena, Italy: 1687 Newton, *Principia mathematica*

MICHIEL HAZEWINKEL Centre of Mathematics and Informatics, Amsterdam, The Netherlands: 1901 Hilbert, paper on problems in mathematics

ALAIN HERREMAN Institute of Mathematical Research (IRMAR), University of Rennes, France: 1934 Seifert and Threlfall, *Topologie*, and 1935 Alexandroff and Hopf, *Topologie*

TIM HORDER Department of Human Anatomy and Genetics, University of Oxford, Britain: 1917 Wentworth Thompson, *On growth and form*

GIORGIO ISRAEL Department of Mathematics, University of Rome, Italy: 1931 Volterra, *Lutte de la vie*

DOUGLAS JESSEPH Department of Philosophy and Religion, North Carolina State University, Raleigh, USA: 1734 Berkeley, *The analyst*

OLE KNUDSEN History of Science Department, Aarhus University, Denmark: 1904 Thomson, *Baltimore lectures*

ANNE KOX Department of Theoretical Physics, University of Amsterdam, The Netherlands: 1909 Lorentz, *Theory of electrons*

JA HYON KU Faculty of Liberal Arts, Youngsan University, Yangsan, South Korea: 1877–1878 Rayleigh, *Theory of sound*

ALBERT LEWIS Peirce Edition Project, Indiana University, Indianapolis, USA: 1844 Grassmann, *Ausdehnungslehre*; and 1853 Hamilton, *Lectures on quaternions*

JESPER LUTZEN Department of Mathematics, Copenhagen University, Denmark: 1894 Hertz, *Prinzipien der Mechanik*

EILEEN MAGNELLO Wellcome Institute for the History of Medicine, University of London, Britain: 1900 Pearson, paper on the chi-squared test

DAVID MASCRÉ Ministry of Foreign Affairs, Paris, France: 1867 Riemann, thesis on trigonometric series

JEAN MAWHIN Mathematical Institute, University of Louvain, Louvain-la-Neuve, Belgium: 1893 Lyapunov, *Stability of motion*

GLEB MIKHAILOV Russian National Committee on Mechanics, Moscow, Russia: 1738 Daniel Bernoulli, *Hydrodynamica*

OLAF NEUMANN Mathematical Institute, Jena University, Germany: 1801 Gauss, *Disquisitiones arithmeticae*

JEAN-PIERRE POTIER Department of Economics, University of Lyon 2, France, and Centre Walras, Lyon: 1871 Jevons, *Theory of political economy*

HELMUT PULTE Institute of Philosophy, Ruhr-University Bochum, Germany: 1788 Lagrange, *Méchanique analitique*

MARK RAVAGLIA Department of Philosophy, Carnegie-Mellon University, Pittsburgh, USA: 1934, 1939 Hilbert and Bernays, *Grundlagen der Mathematik*

KARIN REICH Institute for the History of Science, University Hamburg, Germany: 1748 Euler, *Introductio*; and 1847 von Staudt, *Geometrie der Lage*

HELMUT RECHENBERG Max-Planck-Institute for Physics, Munich, Germany: 1930 Dirac, *Quantum mechanics* and 1932 von Neumann, *Quantenmechanik*

SILVIA ROERO Department of Mathematics, University of Turin, Italy: 1684–1693 Leibniz, first three calculus papers

ERIK LARS SAGENG St. John's College, Annapolis, Maryland, USA: 1742 MacLaurin, *Treatise on fluxions*

JOEL SAKAROVITCH Faculty of Mathematics and Informatics, University of Paris 5, France; and School of Architecture, Paris-Malaquais, France: 1795 Monge, *Géométrie descriptive*

TILMAN SAUER Einstein Papers Project, Caltech, Pasadena, California, USA: 1916 Einstein, paper on general relativity theory

NORBERT SCHAPPACHER Faculty of Mathematics, Strasburg University, France: 1897 Hilbert, report on algebraic number fields

KARL-HEINZ SCHLOTE Saxon Academy of Science at Leipzig, Germany: 1930–1931 van der Waerden, *Moderne Algebra*

IVO SCHNEIDER History of Science, University of the German Armed Forces, Munich, Munich, Germany: 1713 James Bernoulli, *Ars conjectandi* and 1718 De Moivre, *Doctrine of chances*

MICHEL SERFATI IREM, Department of Mathematics, University of Paris 7, France: 1649 Descartes, *Geometria*

WILFRIED SIEG Department of Philosophy, Carnegie-Mellon University, Pittsburgh, USA: 1934, 1939 Hilbert and Bernays, *Grundlagen der Mathematik*

DAVID SINGMASTER Faculty of Mathematics, South Bank University, London, Britain: 1892 Rouse Ball, *Mathematical recreations*

FRANK SMITHIES late of Jesus College, Cambridge, Britain: 1825, 1827 Cauchy, paper and booklet on complex-variable analysis

JACQUELINE STEDALL The Queen's College, Oxford, Britain: 1656 Wallis, *Arithmetica infinitorum*

STEPHEN M. STIGLER Department of Statistics, University of Chicago, USA: 1812 Laplace, *Théorie analytique des probabilités* and 1814 *Essai philosophique*

MICHAEL TOEPELL Faculty of Education, Leipzig University, Germany: 1899 Hilbert, *Grundlagen der Geometrie*

PETER ULLRICH Department of Mathematics, Siegen University, Germany: 1851 Riemann, thesis on complex analysis

JAN VAN DAAL Centre Walras, Lyon, France: 1871 Jevons, *Theory of political economy*

JAN VON PLATO Department of Philosophy, University of Helsinki, Finland: 1933 Kolmogorov, *Wahrscheinlichkeitsrechnung*

CURTIS WILSON St. John's College, Annapolis, Maryland, USA: 1809 Gauss, *Theoria motus*

NORTON WISE Department of History, University of California at Los Angeles, USA: 1867 Thomson and Tait, *Treatise on natural philosophy*

IDO YAVETZ Cohn Institute for History and Philosophy of Science, Tel-Aviv University, Israel: 1892 Heaviside, *Electrical papers*

JOELLA YODER Independent scholar, Newcastle, Washington, USA: 1673 Huygens, *Horologium oscillatorium*

RICHARD ZACH Department of Philosophy, University of Calgary, Canada: 1931 Gödel, paper on incompleteness theorems