John W. Dawson, Jr.

Why Prove it Again?

Alternative Proofs in Mathematical Practice

Birkhäuser

John W. Dawson, Jr.

Why Prove it Again?

Alternative Proofs in Mathematical Practice

with the assistance of Bruce S. Babcock
and with a chapter by Steven H. Weintraub

Birkhäuser

John W. Dawson, Jr.
Penn State York
York, PA, USA

ISBN 978-3-319-17367-2 ISBN 978-3-319-17368-9 (eBook)
DOI 10.1007/978-3-319-17368-9

Library of Congress Control Number: 2015936605

Mathematics Subject Classification (2010): 00A35, 00A30, 01A05, 03A05, 03F99

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

To Solomon Feferman,
friend and mentor,
who suggested I write this book

Preface

This book is an elaboration of themes that I previously explored in my paper "Why do mathematicians re-prove theorems?" (Dawson 2006). It addresses two basic questions concerning mathematical practice:

1. What rationales are there for presenting new proofs of previously established mathematical results?
   and
2. How do mathematicians judge whether two proofs of a given result are essentially different?

The discovery and presentation of new proofs of results already proven by other means has been a salient feature of mathematical practice since ancient times.[1] Yet historians and philosophers of mathematics have paid surprisingly little attention to that phenomenon, and mathematical logicians have so far made little progress in developing formal criteria for distinguishing different proofs from one another, or for recognizing when proofs are substantially the same.

A number of books and papers have compared alternative proofs of particular theorems (see the references in succeeding chapters), but no extended general study of the roles of alternative proofs in mathematical practice seems hitherto to have been undertaken.

Consideration of particular case studies is, of course, a necessary prerequisite for formulating more general conclusions, and that course will be followed here as well. The aim is *not*, however, to arrive at any *formal* framework for analyzing differences among proofs. It is rather

a) to suggest some *pragmatic criteria* for distinguishing among proofs, and
b) to enumerate reasons *why* new proofs of previously established results have so long played a prominent and esteemed role in mathematical practice.

[1] Wilbur Knorr, e.g., noted that "multiple proofs were frequently characteristic of pre-Euclidean studies" (Knorr 1975, p. 9).

Chapter 1 addresses the first of those aims, following clarification of some pertinent logical issues. Chapter 2 then outlines various purposes that alternative proofs may serve. The remaining chapters provide detailed case studies of alternative proofs of particular theorems. The different proofs considered therein both illustrate the motives for giving alternative proofs and serve as benchmarks for evaluating the worth of the pragmatic criteria in terms of which they are analyzed.

York, PA, USA John W. Dawson, Jr.

References

Dawson, J.: Why do mathematicians re-prove theorems? Philosophia Mathematica (III) **14**, pp. 269–286 (2006)

Knorr, W.: The Evolution of the Euclidean Elements. A Study of the Theory of Incommensurable Magnitudes and its Significance for Early Greek Geometry. Reidel, Dordrecht (1975)

Acknowledgments

I am indebted to Solomon Feferman, Akihiro Kanamori, and two anonymous reviewers for suggesting improvements to earlier versions of chapters 1–8. I thank the participants in the Philadelphia Area Seminar on the History of Mathematics, as well as Andrew Arana and Jeremy Avigad, for their enthusiasm for and encouragement of this endeavor. Above all, I am grateful to Bruce Babcock, for preparing the illustrations throughout this book and for his careful copy-editing, and to Steven H. Weintraub, for several enhancements to chapter 6 and for contributing chapter 11 on proofs of the irreducibility of the cyclotomic polynomials. Finally, I thank Anthony Charles and the rest of the Birkhäuser production staff for their cordial and efficient efforts to transform my manuscript into print.

Contents

Chapter 1
Proofs in Mathematical Practice

Before proceeding to consider the questions posed in the Preface, it is necessary to clarify some logical issues. Paramount among them is the question: **What is a proof?**

The notion of what constitutes a proof in a *formalized* theory is perfectly precise: It is a finite sequence of well-formed formulas, the last of which is the statement to be proved and each of which is either an axiom, a hypothesis, or the result of applying one of a specified list of rules of inference to previous formulas of the sequence. That notion of proof, central to mathematical logic, has led to important advances in the understanding of many foundational and metamathematical issues, and it is widely believed that all current mathematical theories can be formalized within the framework of first-order Zermelo-Fraenkel set theory. Nevertheless, formal proofs are not the focus of the present inquiry.

One reason they are not is that formal proofs have not yet become the stuff of mathematical practice:[1] The notion of a formalized theory is of very recent origin, and strictly formal proofs appear almost exclusively in texts on logic and computer science, not in ordinary mathematical discourse. In addition, mathematicians do not ordinarily resort to formalization in order to judge whether two (informal) proofs that deduce the same conclusion from the same premises are essentially the same. Rather, it is usually easy to tell, on informal grounds, whether such proofs are essentially different or merely variants of one another. Intuitively, they are different if they employ *different concepts or tactics*.

Furthermore, as Yehuda Rav has stressed, mere *expressibility* within the language of set theory does not imply that "all ... current *conceptual proofs* can be formalized as *derivations*" within that theory (Rav 1999, p. 20, fn 20). That is so in part because informal proofs often involve "topic-specific moves" that "have no independent

[1]That may well change soon, however, given that computer proofs of such major results as the Four-color Theorem, the Prime Number Theorem, and the Jordan Curve Theorem have now been obtained.

J.W. Dawson, Jr., *Why Prove it Again?*, DOI 10.1007/978-3-319-17368-9_1

logical justification," but serve rather as conceptual "bridges between the initially given data, or between some intermediate steps, and subsequent parts of the argument" (*Ibid.*, p. 26). And, as the logician Jon Barwise noted,

> current formal models of proof are severely impoverished .... For example, ... proofs where one establishes one of several cases and then observes that the others follow by symmetry considerations [constitute] a perfectly valid (and ubiquitous) form of mathematical reasoning, but I know of no system of formal deduction that admits of such a general rule. (Barwise 1989, p. 849)

That is not to say, however, that formal methods are of no use in comparing informal proofs. Indeed, formal methods may sometimes help to clarify whether two proofs of a statement really establish the same result. For example, one primary motive for presenting alternative proofs is to eliminate *superfluous hypotheses* (as in Goursat's improvement of Cauchy's theorem on integrals of analytic functions) or *controversial assumptions* (such as the Axiom of Choice). But **should a proof based on fewer or weaker hypotheses be regarded as establishing the *same* or a *stronger* theorem?**

Suppose, for example, that a theorem T is first proved from a set of hypotheses H, but that later a proof of T is found that employs as hypotheses only some proper subset P of H. Have we given two different proofs of T, or have we proved in the first case the implication $\bigwedge H \Rightarrow T$ (where $\bigwedge H$ denotes the conjunction of all the hypotheses in H) and in the second case the implication $\bigwedge P \Rightarrow T$ (a stronger result)? Formally, the distinction is merely a matter of perspective, since the Deduction Theorem for first-order logic establishes the equivalence of $A \vdash T$ and $\vdash \bigwedge A \Rightarrow T$ for any set A of non-logical axioms. To avoid confusion, however, the first perspective will be adopted here, according to which it is the *proof*, rather than the *theorem*, that has been strengthened; and for precision, all and only those premises that are actually employed in a proof will be regarded as the hypotheses of the deduction. It then follows that two proofs of the same theorem based on logically *inequivalent* sets of premises must be regarded as different, since the totality of contexts (models of the premises) in which one proof is valid is not the same as those in which the other is.

In some cases, formal considerations may lead to conclusions that differ from those arrived at informally. For example: **If a statement *S* is known to imply a statement *T*, should a proof of *S ipso facto* be regarded as a proof of *T*?** Informally, the answer ought to be "no" in general, since the proof that S implies T may itself be highly non-trivial.[2] Formally speaking, however, it follows from (the strong form of) Gödel's completeness theorem for first-order logic that if S and T are *any* two theorems provable from some first-order set of axioms A, then so is

[2]Nevertheless, as a colleague has rightly noted, if T is a statement whose proof has long been sought and the implication $S \Rightarrow T$ has already been established, one who proves S is often said to have proved T. For example, Andrew Wiles proved the Taniyama conjecture, but is often said to have proved Fermat's Last Theorem.

$S \Rightarrow T$: For both S and T, and therefore also the equivalence $S \iff T$, must hold in every structure that satisfies the axioms, and so, by Gödel's theorem, $S \iff T$ must be provable from the axioms A. Thus, given a proof of S, applying *modus ponens* to the proofs of S and of $S \Rightarrow T$ would yield a proof of T, even though, from a semantic standpoint, S might be utterly irrelevant to T.

Such proofs do not occur in actual mathematical practice. But what if the implication $S \Rightarrow T$ is more easily seen (especially if S is harder to prove than T)? For example, should a proof that the series of reciprocals of the primes diverges, or a proof of Bertrand's Postulate, be deemed a proof of the infinitude of the primes? Should a proof of the Pythagorean Theorem be deemed a proof of the Law of Cosines? The answer is somewhat subjective, and it seems impossible to draw a clear-cut boundary. In Aigner and Ziegler's *Proofs from the Book* (Aigner and Ziegler 2000), for example, a proof that the aforementioned series diverges is included among proofs of the infinitude of the primes, but a proof of Bertrand's Postulate is given in a separate section. As for the Law of Cosines, a direct proof of it — that is, one that does not employ the Pythagorean Theorem[3] — certainly establishes the Pythagorean Theorem as a special case. But in practice, as in Euclid, the Pythagorean Theorem is proved first and then used as a tool to prove the Law of Cosines — a proof which, though relatively straightforward, is not trivial. Thus, although logically equivalent to the Law of Cosines, in practice the Pythagorean Theorem exhibits a certain *conceptual primacy*. It does not seem proper, therefore, to regard a direct proof of the Pythagorean Theorem as a proof *per se* of the Law of Cosines; rather, one may distinguish proofs of the latter according to whether they do or do not rely on the Pythagorean Theorem.

We do not, then, eschew the use of formal *methods* in the analyses to be undertaken here. But the formal model of what constitutes a proof is an abstraction designed to "provide an explanation of … [how] an informal proof is judged to be *correct* [and] what it means for a [mathematical statement] to be a *deductive consequence*" of certain other statements.[4] For those purposes, the formal model of proof serves very well. It seems ill suited, however, for dealing with the broader sort of questions considered here.

Accordingly, the term 'proof' will here be taken to refer to an informal argument, put forward to convince a certain audience that a particular mathematical statement is true — and, ideally, to explain *why* it is true — an argument that is subsequently accepted as valid by consensus of the mathematical community. As such, whether a proof succeeds in producing conviction that the result it purports to prove is true

[3]It seems that only very recently has such a proof of the Law of Cosines been given. See http://www.cut-the-knot.org/pythagoras/CosLawMolokach.shtml (discussed further in Chapter 5 below).

[4]Quoted from Avigad (2006), an article whose concerns overlap to some extent with those of the present text. Avigad suggests that a more fruitful model for analyzing broader aspects of proofs that occur in mathematical practice may be that employed by workers in the field of automated deduction.

depends not only on the formal *correctness* of the argument, but on the mathematical knowledge and sophistication of the audience to which it is presented, as well as that of the mathematical community at large.

The issue here is both a historical and a pedagogical one: On the one hand, standards of rigor have not remained constant, so arguments that once were accepted as convincing by the community of mathematicians of the time may no longer be so regarded; and who is to say that a proof accepted as valid today will not some day be found wanting?[5] On the other hand, a rigorously correct proof may fail to be convincing to those who lack the requisite background or mathematical maturity; and some results (such as the Jordan Curve Theorem) may appear so obvious that mathematical sophistication is required even to understand the *need* for them to be proved.

Recognition of those facts is essential for any meaningful study of the role of alternative proofs in mathematical practice. For to dismiss as proofs arguments that once were, but are no longer, deemed to be correct or complete is to misrepresent mathematical history, by attributing to proofs a permanence they do not possess; and in the present context, it would also eliminate from consideration two of the primary motives for seeking alternative proofs, namely, *correcting errors or filling perceived gaps in previous proofs* (as, e.g., in Hilbert's rigorization of Euclidean geometry), and *presenting arguments that*, though perhaps less rigorous, *are more perspicuous or persuasive to a given audience*.[6]

The primary aim in what follows will be to examine how alternative proofs of various well-known theorems differ. In most cases it will be evident that the proofs *do* differ, and that intuitive feeling can be justified in various ways. For example, proofs that are *direct*, or are *constructive*, may be distinguished from those that are not. A proof that employs a *particular technique* (mathematical induction, for example, or a certain rule of inference) differs tactically from one that does not. One proof may give *greater information* than another — for example, by providing a method for finding a solution to an equation, rather than merely exhibiting one, or by better indicating *why* a result is true. (Showing, e.g., that a convergent real power series has a particular radius of convergence by showing that the corresponding *complex* series has a singularity at a certain point is more informative than simply

[5]It is interesting to note, however, how many arguments later deemed to be 'faulty' have yielded correct results — have contained a 'germ' of truth, so to speak. In some cases, the methods originally used to prove such results have been discarded and the theorems reestablished by other, quite different means, while in other instances the original approaches have subsequently been revalidated in light of more sophisticated analyses. (One example is Laurent Schwartz's theory of distributions, which provided a rigorous foundation for arguments based on Dirac's 'δ-function.' Another is Abraham Robinson's creation of non-standard analysis, in terms of which the Newtonian concept of infinitesimal was made comprehensible and arguments based upon it were seen to be correct.)

[6]Proofs may, for example, be crafted to serve the needs of a particular segment within or outside of the mathematical community (students, for example, or lay persons with an interest in mathematics).

applying the ratio test to the given series.) Different proofs may yield *different numerical consequences*, one yielding a better numerical bound, say, than another, or a smaller number that exhibits a certain property. A result employed as a *lemma* in one proof of a theorem may appear as a *corollary* of that theorem if it is proved by other means. One proof of a theorem may be valid in a *wider context* than another. Or one proof of a theorem may be *comprehensible to a particular audience* while another is not.

Proofs may also differ in *how primitive notions are organized into higher-level concepts*, reflecting an *interplay* between *proofs* and *definitions*.

Consider, for example, proving that $\det(AB) = \det(A)\det(B)$. If, for $n \times n$ matrices A, $\det(A)$ is defined as $\sum_{\sigma \in S_n} \operatorname{sgn}(\sigma)\, a_{1\sigma_1} \dots a_{n\sigma_n}$, where S_n denotes the set of all permutations of $\{1, \dots, n\}$ and a_{ij} denotes the entry in row i and column j of A, one may compute directly that

$$\begin{aligned}
\det(AB) &= \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \left(\sum_{k_1} a_{1k_1} b_{k_1\sigma_1} \right) \dots \left(\sum_{k_n} a_{nk_n} b_{k_n\sigma_n} \right) \\
&= \sum_{k_1, \dots, k_n} a_{1k_1} \dots a_{nk_n} \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) b_{k_1\sigma_1} \dots b_{k_n\sigma_n} \\
&= \det(B) \sum_{k \in S_n} \operatorname{sgn}(k) a_{1k_1} \dots a_{nk_n} \\
&= \det(B)\det(A) = \det(A)\det(B);
\end{aligned}$$

but the defining formula is complicated and the index chasing even more so.

Instead, with that same definition for $\det(A)$, many linear algebra texts introduce the elementary row operations and the corresponding elementary matrices E, show how each row operation on A results in the matrix EA, and compute easily that $\det(EA) = \det(E)\det(A)$ for any $n \times n$ matrix A and any elementary matrix E. If A is nonsingular, some sequence of m row operations reduces it to the $n \times n$ identity matrix I, so $E_m \dots E_1 A = I$. Noting that the inverse of any row operation is itself a row operation, $A = E_m^{-1} \dots E_1^{-1}$, so

$$\begin{aligned}
\det(AB) &= \det(E_m^{-1} \dots E_1^{-1} B) \\
&= \det(E_m^{-1}) \dots \det(E_1^{-1}) \det(B) \\
&= \left(\det(E_m^{-1}) \dots \det(E_1^{-1})\right) \det(B) \\
&= \det(A)\det(B).
\end{aligned}$$

On the other hand, if A is singular, then so is AB, for any $n \times n$ matrix B, so A and AB each reduce to a matrix with one or more rows of zeros, whence $\det(AB) = 0 = \det(A) = \det(A)\det(B)$.

The proof just given is much more perspicuous than the direct one. However, because of its use of inverses, it presumes that the entries in the matrices are elements of a field, whereas the direct proof applies in the wider context of matrices whose entries are elements of a commutative ring.

A third alternative, adopted in texts such as Hoffman and Kunze (1961), is to define a determinant function to be a function from $n \times n$ matrices over a commutative ring R to elements of that ring which is linear as a function of each row of the matrix, which assigns the value 0_R to any matrix having two identical rows, and which assigns the value 1_R to the identity matrix. One then proves that there is exactly one such function, which may be expressed in terms of the entries of the matrix by the aforementioned formula. Without reference to that formula, however, one can show that if D is any function from the $n \times n$ matrices over R to R that is a linear function of each row of any such matrix and that is *alternating* (that is, $D(A) = -D(A')$ whenever A' is obtained from A by interchanging two of its rows), then $D(A) = \det(A)D(I)$. Such a function D is given by $D(A) = \det(AB)$, for *any* fixed $n \times n$ matrix B. So

$$\det(AB) = D(A) = \det(A)D(I) = \det(A)\det(IB) = \det(A)\det(B).$$

In this example, changing the definition of $\det(A)$ from one couched in terms of the entries of the matrix to one that incorporates some of the desired *properties* of the determinant function leads, without loss of generality, to a more abstract proof that is both more perspicuous than the computational one and that explains *why* the formula for $\det(A)$ in terms of the entries of A is what it is.

In comparing different proofs, it must of course be recognized that differences are often a matter of degree. The judgment whether two proofs are essentially different or merely variants of one another (and similar judgments as to whether one is 'simpler', or 'more pure', than another — notions discussed further in the next chapter) is thus a subjective one to a certain extent. Chapter 3 provides a very simple but illustrative example.

The next chapter enumerates some of the reasons why mathematicians have been led to seek alternative proofs of known results. The remaining chapters then provide detailed case studies of particular instances.

References

Aigner, M., Ziegler, G.M.: Proofs from the Book, 2nd ed. Springer, Berlin (2000)
Avigad, J.: Mathematical method and proof. Synthese **193**(1), 105–159 (2006)
Barwise, J.: Mathematical proofs of computer system correctness. Notices Amer. Math. Soc. **36**(7), 844–851 (1989)
Hoffman, K., Kunze, R.: Linear Algebra. Prentice-Hall, Englewood Cliffs, N.J. (1961)
Rav, Y.: Why do we prove theorems? Philosophia Math.(III) **7**, 5–41 (1999)

Chapter 2
Motives for Finding Alternative Proofs

> *Even if we have succeeded in finding a satisfactory solution, we may still be interested in finding another solution. We desire to convince ourselves of the validity of a theoretical result by two different derivations as we desire to perceive a material object through two different senses. Having found a proof, we wish to find another proof as we wish to touch an object after having seen it.*
>
> — George Pólya, *How to Solve It*

Four motives for seeking new proofs of previously established results have already been mentioned in Chapter 1: the desires

(1) **to correct errors or fill perceived gaps in earlier arguments;**
(2) **to eliminate superfluous or controversial hypotheses;**
(3) **to extend a theorem's range of validity; and**
(4) **to make proofs more perspicuous**.

Euclid's efforts to avoid, wherever possible, employing proofs involving superposition of figures exemplify the first of those motives; the persistent attempts, prior to the works of Bolyai and Lobachevsky, to deduce the parallel postulate from the other axioms of Euclidean geometry, the second; Henkin's completeness proof for first-order logic, applicable to uncountable languages as well as to countable ones, the third; and objections, by sixteenth- and seventeenth-century mathematicians, to Archimedean proofs by the method of exhaustion, the fourth.

That Euclid employed the method of superposition when it appeared unavoidable, as in the proof of proposition I,4 (justifying the side-angle-side criterion for congruence), suggests that the ancients considered superposition to be a *perspicuous* principle, but one that rested on *spatial* intuition and so was not rigorously justified by Euclid's (planar) axioms. On the other hand, the *rigor* of Archimedes's proofs by exhaustion was never questioned. But such proofs seemed to many merely to establish *that* a result was true, without providing understanding of *why* it was, or of how the proof might have been discovered.[1] Proofs by mathematical induction are open to similar objections.

[1] There is a growing literature on the notion of *explanatory* proofs (those that convey understanding as well as conviction). The article Mancosu (2001) and the book Mancosu (1996) provide useful introductions to that subject.

J.W. Dawson, Jr., *Why Prove it Again?*, DOI 10.1007/978-3-319-17368-9_2

The desire to avoid employing superposition arguments in Euclidean proofs also exemplifies another respect, separate from that of rigor, in which a proof may be deemed deficient: that it fails to exhibit **purity of method**.

Concern for purity of method has arisen frequently in the history of mathematics. Particular instances of such concern include the ancient Greek requirement that geometric constructions be restricted to those performable with straightedge and compass alone; the desires that synthetic proofs be found for results obtained in analytic geometry, that intrinsic proofs be given for theorems in differential geometry or topology, and that proofs of results in model theory not invoke syntactic considerations; the preference for minimizing appeals to analytical or topological methods in proving the Fundamental Theorem of Algebra (discussed further in Chapter 8 below); the quest to find an 'elementary' proof of the Prime Number Theorem[2] (one not employing the methods of analytic number theory), whose unexpected success was among the achievements that led to the award of a Fields Medal to Atle Selberg; and Hilbert's (failed) program to establish the consistency of formalized Peano arithmetic using methods formalizable within that theory itself. In a somewhat broader sense, **concern for methodological propriety** is reflected in such aspects of mathematical practice as the desire to replace indirect or non-constructive proofs by direct or constructive ones, or the debate over whether theorems of analysis ought to be proved by 'soft' (functional-analytic) means (which, though 'slick,' may obscure underlying conceptual motivations) or by 'hard' calculations involving inequalities (which may be lengthy and tedious).

In considering questions of purity, several caveats are in order. First, as remarked in the preceding chapter, in many cases one should more properly speak of *degrees* of purity. Second, it must be recognized that different notions of purity may *conflict*. Consider, for example, Desargues's Theorem in the Plane, which states that if two triangles in the same plane are oriented so that the lines joining corresponding vertices are concurrent, then the corresponding sides, if extended as lines, will intersect in three collinear points. As a theorem of *projective* geometry, one might in the name of purity seek a proof solely by projective means. But as a theorem of *plane* geometry, one might equally well desire a purely planar proof. Those two aims cannot be reconciled, however, since Hilbert showed that any planar proof of that theorem must invoke the *metric* notion of similarity.[3] In addition, it should be noted that some proofs — for example, model-theoretic consistency proofs, in which basic notions are given alternative semantic interpretations — inherently violate purity of method.

Proofs that violate purity of method may, however, possess merits of their own. The example given in the previous chapter of determining the radius of convergence

[2] In its simplest form, the Prime Number Theorem (the subject of Chapter 10 below) states that $\lim_{x\to\infty} \pi(x) \Big/ \dfrac{x}{\ln x} = 1$, where $\pi(x)$ denotes the number of primes less than x.

[3] See Chapter 9 below for a detailed discussion of Desargues's Theorem. An illuminating discussion of Hilbert's proof is given on pp. 222–229 of Hallett (2008).

of a real power series is a case in point: consideration of the singularities of the corresponding complex series makes it clear *why* the radius of convergence is what it is — an insight that could not be obtained without introducing the complex perspective. The more general context here has greater explanatory power.

Concern for purity of method implies a restriction on means of proof, but not conversely. Other reasons for restricting allowable means of proof include **reconstructing proofs employed in antiquity**, given what we know about the state of mathematical knowledge in particular ancient cultures,[4] and **benchmarking** (demonstrating the power of a given methodology by employing it to prove theorems in areas where it might seem not to be applicable).[5]

Methodologically 'pure' proofs demand fewer conceptual prerequisites for their understanding, since they employ no notions beyond those implicit in the statement of the theorem to be proved. They may, however, be long and complex (as are elementary proofs of the Prime Number Theorem), and thus lack **elegance**, an aesthetic characteristic that is hard to define but is nonetheless readily perceived and highly esteemed by mathematicians.

Proofs that are elegant may employ sophisticated concepts, but they are usually short, often employ novel perspectives or strategies, and generally convey immediate understanding and conviction (producing an *Aha!* reaction). Reading such proofs yields deep intellectual enjoyment and satisfaction, akin to that experienced in viewing fine works of art or listening to great music. Elegant proofs are, however, often of limited generality, involving insights applicable only to a particular problem.

Another aesthetic criterion according to which proofs may be compared is **simplicity**. One proof may be simpler than another in various respects. For example:

(1) It may be significantly shorter.
(2) It may involve fewer conceptual prerequisites.
(3) It may reduce the extent of computations to be performed or the number of cases to be considered.

A well-known example of (3) is Hilbert's basis theorem for invariants of algebraic forms,[6] whose non-constructive proof swept away in one stroke a tangle of laborious calculations in invariant theory, including much of the life work of the mathematician Paul Gordan.

[4] One example is discussed in Chapter 4.

[5] Such as using topological arguments to prove results in mathematical logic. Another example is Errett Bishop's text *Foundations of Constructive Analysis* (Bishop 1967), which Bishop himself called "a piece of constructivist propaganda," written to demonstrate how large a part of abstract analysis can be developed within a constructive framework.

[6] In the form proved by Hilbert, the theorem states that every ideal in the ring of multivariate polynomials over a field is finitely generated, so that for any set of polynomial equations, there is a finite set of such equations that has the same set of solutions.

Remarkably, Hilbert himself believed that among all proofs of a theorem there must always be one that is simp*lest*. He said so explicitly in the statement of what he had intended would be the twenty-fourth problem in the list he drew up for presentation in his famous address at the Second International Congress of Mathematicians. Due to time constraints, however, he mentioned only ten of the problems during the lecture itself. Thirteen more appeared in the version of his address published in the Conference *Proceedings*, but the twenty-fourth problem came to light only in the mid-1990s, when it was discovered in one of the notebooks in Hilbert's *Nachlass*.[7] It asked for "Criteria of simplicity, or proof of the greatest simplicity of certain proofs," with the understanding that "under a given set of conditions there can be but one simplest proof." Hilbert never posed the problem in public, perhaps because of the difficulty of making the notion of "simplicity" formally precise.[8] Accordingly, judgments of whether one proof is simpler than another have up to now been based primarily on informal criteria like those above.

In addition to the practical and aesthetic rationales so far considered for presenting new proofs of previously established theorems, there are also more personal motives for doing so. For mathematics is, after all, a *human* endeavor. Skill in proving theorems is best developed by attempting to prove results on one's own, and in the course of doing so, one may well devise an argument not previously given, since people do not all think alike. In some cases, one may not be *aware* of other proofs that have been given—different proofs, e.g., may arise in different cultures, or a new result may be discovered, simultaneously and independently, by different individuals using different arguments. Like all fields of scholarship, mathematics is also a competitive enterprise, and having seen a proof presented by someone else, one may be challenged to devise one's own proof of it. Thus, apart from the reasons already enumerated, alternative proofs may arise simply as **expressions of individual patterns of thought**, perhaps reflecting personal predilections or preferences for using particular tools.

Here an analogy may be made between mathematics and the sport of mountaineering. Mathematicians are driven to solve problems for the same reason that mountaineers are driven to climb mountains: because they are there; and as in mountaineering, **pioneering a new route** to a summit, perhaps using restricted means, may be just as challenging and exciting (and be accorded just as much respect by one's peers) as being the first to make the ascent.[9] Mascheroni's work showing that the compass alone suffices to carry out all straightedge and compass constructions may, for example, be compared with ascents by climbers who disdain

[7] "Mathematisches Notizbuch" (Cod. ms. D. Hilbert 600), preserved in the Handschriftenabteilung of the Niedersächische Staats- und Universitätsbibliothek, Göttingen.

[8] See Thiele and Wos (2002) for further details on the history of the twenty-fourth problem and on results related to it found recently by those working in automated theorem proving.

[9] Jon Krakauer, e.g., in his book *Into Thin Air*, wrote: "Getting to the top of any given mountain was considered much less important than *how* one got there: prestige was earned by tackling the most unforgiving routes with minimal equipment, in the boldest style imaginable."