Philosophia Mathematica (III) **14** (2006), 269–286. doi:10.1093/philmat/nkl009
Advance Access publication June 9, 2006

Why Do Mathematicians Re-prove Theorems?[†]

JOHN W. DAWSON, JR*

From ancient times to the present, the discovery and presentation of new proofs of previously established theorems has been a salient feature of mathematical practice. Why? What purposes are served by such endeavors? And how do mathematicians judge whether two proofs of the same theorem are essentially different? Consideration of such questions illuminates the roles that proofs play in the validation and communication of mathematical knowledge and raises issues that have yet to be resolved by mathematical logicians. The Appendix, in which several proofs of the Fundamental Theorem of Arithmetic are compared, provides a miniature case study.

The discovery and proof of new *theorems* is presumed by many to be the *sine qua non* of mathematical endeavor. Yet since ancient times the presentation of new *proofs* of previously established results has also been an esteemed and commonplace mathematical practice. Wilbur Knorr, for example, in his book *The Evolution of the Euclidean Elements*, noted that 'multiple proofs were frequently characteristic of pre-Euclidean studies', and he went on to show that consideration of alternative proofs could elucidate some historical puzzles concerning ancient mathematics ([Knorr, 1975]; discussed further below). Today, new proofs of old theorems continue to appear regularly and to enrich mathematics. Indeed, in 1950 a Fields Medal was awarded to Atle Selberg, in part for his elementary proof of the prime-number theorem. So one is led to ask: What reasons are there for re-proving known results? And how do mathematicians judge whether a proof is conceptually distinct from those that have been given before?

The efforts mathematicians have devoted to re-proving theorems are understandable if, as Yehuda Rav has contended in his provocative paper [Rav, 1999], proofs themselves are the primary focus of mathematical concern. Rav does not adduce the re-proving of theorems as evidence for that thesis, but many of the issues he raises are central to the discussion that follows. In particular, like him, we must confront at the outset the question 'What constitutes a proof?'.

[†] I wish to thank Andrew Arana, Solomon Feferman, Mary Leng, Paolo Mancosu, and Yehuda Rav for suggesting improvements to the original draft of this paper.

* Penn State York, 1031 Edgecomb Ave., York, Pennsylvania 17403 U. S. A. jwd7too@suscom.net

1. Formal versus Informal Proofs

For *formalized* theories, the notion of a proof is completely precise: It is a sequence of well-formed formulas, the last of which is the theorem that is proved and each of which is either an axiom or the result of applying a rule of inference to previous formulas in the sequence. We shall *not*, however, adopt that definition. Rather, we shall take a proof to be an *informal* argument whose purpose is to convince those who endeavor to follow it that a certain mathematical statement is true (and, ideally, to explain *why* it is true).

There are several reasons why the formal notion of proof is not appropriate in the present context. First of all, formal proofs are artificial constructs of very recent origin. They are abstractions from mathematical practice that fail to capture many important aspects of that practice. The late Jon Barwise, one of the world's most distinguished logicians, stressed that 'the idea that reasoning [can] ... somehow be reduced to syntactic form in a formal ... language' originated with Hilbert early in the twentieth century. And while Hilbert's formalized *proof theory* has proved to be very fruitful in logical investigations, Barwise acknowledged that 'current formal models of proof are severely impoverished'; none of those so far proposed, for example, admit such 'perfectly valid (and ubiquitous) form[s] of mathematical reasoning' as a proof by cases in which one case is established in detail and the others are observed to follow by symmetry considerations [Barwise, 1989].[1]

Restricting attention to formal proofs would exclude most actual proofs from consideration, since few mathematical theories have been formally axiomatized, and many informal proofs do not proceed from explicitly stated assumptions.[2] Moreover, as Rav stresses, while it is generally accepted that most 'current mathematical theories can be *expressed* in first-order set-theoretical language', it does *not* follow that 'all ... current *conceptual proofs* can be *formalized* as *derivations*'([Rav, 1999, p. 20, fn 20]; his emphasis); rather, informal proofs often involve 'topic-specific moves' that 'have no independent logical justification' but serve as conceptual 'bridges between the initially given data, or between some intermediate steps, and subsequent parts of the argument' (*ibid.*, p. 26).[3]

[1] He contended that such considerations *cannot* be incorporated within any formal model of proofs, because they are 'not, in general, something one can determine from local, syntactic features of a proof'. Recent work in proof theory, however, has introduced more global methods of analysis (*e.g.*, the notion of logical flow graph introduced in [Buss, 1991]; *cf.* the review [Carbone, 1997]).

[2] Nor need they even be verbal. The Mathematical Association of America, for example, has published two volumes of *Proofs Without Words.*

[3] Rav cites the introduction of the number $N = p_1 p_2 \ldots p_n + 1$ in Euclid's proof of the infinitude of the primes as one example of such a move.

Furthermore, formalizing mathematical theories in a first-order framework is usually a Procrustean act: [Shapiro, 1991] gives persuasive arguments that second- or higher-order logic is the natural vehicle for formalizing many theories that occur in mathematical practice.

Formal proofs appear almost exclusively in works on computer science or mathematical logic, primarily as objects of study to which other, informal, arguments are applied. If written out in full they are difficult to comprehend, and despite their rigor they are often *unconvincing,* because although they provide *verification* that a result follows logically from given premises, they may fail to convey *understanding* of *why* it does. How else, for example, are we to explain the debate over the legitimacy of computer-assisted proofs, such as that of the Four-color Theorem? Surely, computer-assisted proofs are not more likely to be *mistaken* than humanly generated proofs. Programming errors or hardware failures may of course occur, and noise may corrupt the running of a program; but the performance of the underlying algorithms may be checked by having them programmed independently by others, in other languages, and run on different machines. So the process of validation for computer-assisted proofs differs little from the vetting of humanly generated proofs by other mathematicians, except that it is carried out by machines. Nevertheless, because computer-assisted proofs are *not humanly surveyable*, many find them unsatisfying.

The issue here is not logical, but *psycho*logical: At some point, for example, despite our belief in the transitivity of logical implication, we lose track of conceptual threads when we are presented with a sufficiently long chain of formal deductions.[4] In informal humanly generated proofs, *lemmas* are used to break such chains up into manageable pieces. They serve as signposts to mark important conceptual steps in the proof.

It may still be objected, however, that taking proofs to be 'convincing arguments' rather than formal derivations is too vague and subjective a notion; for whether an argument produces conviction that a mathematical result is true depends not only on its correctness, but on the mathematical sophistication of the audience to whom it is presented.

That, of course, is true, both historically and pedagogically: Because standards of rigor have not remained constant, arguments that once were accepted as convincing may no longer be, while 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 it requires mathematical sophistication even to understand the *need* for a proof.)

[4] A referee has remarked that it is not just the computational complexity of the four-color proof that makes it dissatisfying. Beyond that, 'we would like there to be a simple reason why the theorem holds in all cases.' But it would be naïve to expect that there *must* be such a reason.

Should we then disallow proofs that are not *now* deemed to be rigorous, or those that are rigorous but not convincing?[5]

To dismiss arguments that, years later, are recognized to be incorrect or incomplete is to misrepresent mathematical history by attributing to proofs a permanence they do not possess. Who is to say, for example, that a proof now accepted as valid will not some day be found wanting? And if faulty arguments have no validity, why is it that so many of them turn out to be reparable—to contain, so to speak, a 'germ' of truth? The definition of an informal proof as a convincing argument, held to be so by consensus of the mathematical community at a given time, entails that a proof may not be valid for all time—a point of view which, though at odds with the formal conception, is the only one that seems historically tenable. And in the context of the present paper, the perception that an argument contains gaps or is not convincing to the audience to whom it is presented are two important reasons why mathematicians may seek to find other proofs.

2. Criteria for Differentiating Among Proofs

In practice, mathematicians usually have little difficulty recognizing when one proof is essentially different from another. But what are their criteria for doing so? Or, to turn the question around: Under what circumstances are two proofs of the same theorem, based on the same premises, to be regarded as *conceptually equivalent*?

In the case of *formal* proofs, logicians have yet to agree on what the criteria should be. One proposal, based on the notion of reduction to normal form in Gentzen's natural-deduction calculus, was given by Dag Prawitz [1971]. He posited that two proofs should be regarded as conceptually equivalent if and only if both reduce to the same normal form. The 'if' direction of that assertion seemed 'a reasonable thesis' to him, but Solomon Feferman, in his review [1975] of Prawitz's paper, disagreed. He pointed out that if an atomic formula $A(t)$ of arithmetic is derived via $\forall$-elimination from a formula $\forall \mathrm{x}\, A(\mathrm{x})$ that was itself obtained by $\forall$-introduction, then the reduced derivation, which contains no such successive introductions and eliminations, merely provides a computational verification of $A(t)$. Every abstract idea in the original derivation may be lost in the reduction (which, he noted, was precisely Hilbert's aim in proposing his reductive program of proof theory).

[5] The Four-color Theorem is once again exemplary, on both counts: Kempe's original proof was found to be flawed after years of acceptance by the mathematical community, while the proof of Appel and Haken, though now accepted as rigorous, was at first met with dubiety because it was computer-assisted.

Another proposal, formulated at about the same time by Joachim Lambek, introduced a category-theoretic notion of *generalization*, according to which 'two derivations have the same generality when every generalization of one of them leads to a generalization of the other, [and] ... the two generalizations have the same assumptions and conclusion'. The survey article [Došen, 2003], from which the preceding quotation is taken, offers the following propositional example: the statement p may be derived from the disjunction $p \vee p$ by projecting either on the first disjunct or the second. The first derivation generalizes to the derivation of p from $p \vee q$, while the second generalizes to the derivation of q from $p \vee q$. Since the conclusions of the two generalizations are not the same, the two projections do not have the same generality. Lambek suggested that two derivations should be regarded as representing the same proof if and only if they have the same generality.

Unfortunately, as Došen points out (p. 487), Prawitz's and Lambek's proposals agree 'only for limited fragments of logic'. The former 'fares rather well in intuitionistic logic' (the context in which it was first formulated), but 'not so well in classical logic'. On the other hand, Prawitz's proposal 'applies also to predicate logic', whereas Lambek's 'has not yet been investigated outside propositional logic' (*ibid.*, pp. 492–493). As for proofs outside of pure logic, such as proofs of the Pythagorean Theorem, Došen concedes that it may seem 'hopeless to try to decide', on the basis of either of the two proposals, 'whether they are identical' (*ibid.*, p. 500).

Of course, *different* proofs of the same theorem need not start from the same hypotheses, and from a logical standpoint, that raises the question: Where does the boundary lie between different *proofs* and different *theorems*? Suppose, for example, that a statement T is originally proved using the axioms of Zermelo-Fraenkel set theory (ZF) plus the Axiom of Choice (AC). Later, a proof of T is discovered that does not invoke the Axiom of Choice. Using the symbol $\vdash$ to denote the provability relation between premises and conclusion, we may describe the situation in either of two ways:

(2.1) that it was first established that $\mathrm{ZF} + \mathrm{AC} \vdash T$ and later that $\mathrm{ZF} \vdash T$, *or*

(2.2) that the first proof established that $\mathrm{ZF} \vdash \mathrm{AC} \rightarrow T$ and the second that $\mathrm{ZF} \vdash T$.

By the Deduction Theorem, the two descriptions are equivalent; but in (1) the *hypotheses* are regarded as having been changed (made 'weaker'), and in (2) the *theorem* (made 'stronger'). To avoid ambiguity, we shall adopt the first perspective, and, for precision, regard as premises all and only those premises that are actually used in a deduction. With that understanding, two proofs of the same theorem that are based on logically inequivalent sets of premises must automatically be regarded as different, since the totality of

contexts in which the first proof holds (the set of all models of its premises) does not coincide with the totality of contexts in which the second holds.

On the other hand, given any set A of first-order axioms, if $A \vdash S$ and $A \vdash T$, then by Gödel's Completeness Theorem, $A \vdash S \leftrightarrow T$ (since Gödel's theorem says that a statement is provable from A if and only if it is true in all models of A, and $S \leftrightarrow T$ will hold in all models of A if both S and T do.) That is, in classical first-order logic, *any* two theorems, proved from the *same* set of axioms, are equivalent in all models of those axioms. From a syntactic standpoint, that means that given a proof of S, together with a proof from the same axioms of *any* other statement T, we can obtain another formal proof of S by applying *modus ponens* to the concatenation of proofs of T and of $T \rightarrow S$, even though T may be semantically irrelevant to S.[6] Such arguments, however, do not occur in actual mathematical practice.

How, then, can we recognize conceptual differences among proofs that deduce the same conclusion from logically equivalent sets of premises? There are several possibilities. There may, for example, be *structural differences* in the arguments. The proofs may employ *different strategies or techniques* (one say, using mathematical induction, another not) or *differ in how inference rules are applied* (invoking different rules, applying the rules to different statements, or using the rules in a different order).[7] A result employed as a *lemma* in one proof of a theorem may appear as a *corollary* to the theorem when it is proved another way.[8] One proof may yield greater information than another (for example, by providing a constructive procedure for producing an object, as opposed to a non-constructive proof of its existence). Or the primitive notions involved in the proof may be *organized into concepts in different ways*. Indeed, as Feferman has stressed, advances in mathematical research (and pedagogy, one might add) often result from 'finding suitable abstract concepts around which to wind large parts of [a] subject in an understandable way' [Feferman, 1978].

Intuitively, proofs are regarded as different if they incorporate *different ideas* or *tactics*. But how can such differences be described concretely? One direction that seems worth exploring is akin to George Pólya's notion of *proof heuristics* (see his [1954]). Pólya's primary concern was to elucidate the path to the *discovery* of mathematical results, but similar considerations may be applicable in the context of *justification* as well. Such an approach

[6] It is this 'defect' of classical logic that relevance logics aim to address.

[7] A similar phenomenon, 'constructional homonymity', occurs in transformational linguistics, where the same surface structure may result from different underlying deep structures; in that case, however, the derived sentence exhibits semantic ambiguity.

[8] For example, the Compactness Theorem for (countable) first-order theories, now usually proved as a corollary to Gödel's Completeness Theorem, was a lemma in Gödel's original proof of the latter. See also the Appendix below.

is best illustrated by example. Toward that end, the Appendix is devoted to an examination of how several proofs of the Fundamental Theorem of Arithmetic may be distinguished in informal, 'heuristic', terms.

3. Reasons for Re-proving Theorems

Let us now address the first of the questions posed in the opening paragraph. There are, in fact, many reasons why mathematicians seek alternative proofs of known results. Among them are the following:

(3.1) *To remedy perceived gaps or deficiencies in earlier arguments.*

In some instances, previous proofs are deemed to be so defective that the result is no longer regarded as a theorem; the Four-color Conjecture, once again, may be cited as an example. But such instances are the exception. More often, the result is accepted as correct, but more rigorous or more satisfying demonstrations are sought. A well-known example is Hilbert's rigorization of Euclidean geometry. Certainly the validity of Euclid's results was never in question; but it was gradually realized that certain assumptions, such as those concerning the relation of *betweenness* among points, had tacitly been made in the proofs. During the sixteenth and seventeenth centuries there was also dissatisfaction with many Archimedean proofs (in particular, those employing the method of exhaustion), not on the grounds of rigor, but because they were extremely tedious and *gave no hint how the results had been found*.[9] A detailed discussion of the matter has been given by Paolo Mancosu in his [1996]. To paraphrase his analysis, it was the disjunction between *methods of proof* and *methods of discovery* that was troubling. He notes that the distinction between arguments that merely compel assent and those that also enlighten the mind goes back to Aristotle, and he quotes the opinion of G. Nardi, a mathematician associated with Cavalieri, that 'everything evident is certain but not everything certain is evident' (*ibid*., p. 63).[10]

Particular instances of (3.1) are the desires

(a) to replace a non-constructive argument by a constructive demonstration;
(b) to provide a more efficient algorithm for performing a calculation, as in the recent proof that primality testing can be done in polynomial time; and

[9] Similar objections may be lodged against many proofs by induction in arithmetic: though efficient, they merely verify the correctness of results discovered by other means.

[10] There is now a growing literature on 'explanatory' proofs—those that demonstrate *why* a theorem is true, not just *that* it is true. For an introduction to that topic, see for example [Mancosu, 2001] or [Mancosu, Jørgensen, and Pedersen, 2005].

(c) to eliminate controversial hypotheses, such as the parallel postulate, the Axiom of Choice, the Riemann Hypothesis, or the Continuum Hypothesis.

An example illustrating both (3.1c) and (3.1a) is the proofs, first by James Ax and Simon Kochen and later by Paul Cohen, that the validity of statements concerning the field of p-adic numbers is effectively decidable: Ax and Kochen established the result *in principle* by deducing it on the assumption of the Continuum Hypothesis and then eliminating that hypothesis by appeal to a general conservation result due to Georg Kreisel. Cohen, in contrast, described an explicit decision procedure.

Note, however, that proofs satisfying criterion (3.1a) or (3.1c) are often more intricate than ones they replace; so they do not necessarily produce greater insight as to why a result is true or how it was discovered. And, in contrast to (3.1a), there are many instances in which constructive proofs have *preceded* non-constructive ones. One example is Liouville's explicit construction of a transcendental number, which antedated Cantor's proof (via a cardinality argument) that transcendental numbers must exist. In some cases, the non-constructive proof may be deemed

(3.2) *to employ reasoning that is simpler, or more perspicuous, than earlier proofs.*

A proof may be simpler than others in various (possibly overlapping) ways. For example:

(a) it may *reduce the extent of computations* or the *number of cases* to be considered;
(b) it may be significantly *shorter*;
(c) it may exhibit *economy of means*, for example by employing *fewer hypotheses* (as did Goursat's proof of the Cauchy integral theorem, which dispensed with the hypothesis that the derivative of the integrand be continuous) or by requiring *fewer conceptual prerequisites* (thereby making it comprehensible to a wider audience).

One important reason for seeking a proof based on fewer hypotheses is that a proof based on minimal assumptions is more likely to generalize to other contexts. The 'reverse mathematics' program undertaken by Stephen Simpson and Harvey Friedman, whose aim is to show how much can be proved within a weak subsystem of arithmetic, exemplifies that motivation.

A prime example of (3.2a)—but one running counter to (3.1a)—is Hilbert's basis theorem for invariants of algebraic forms: a non-constructive proof that in one stroke swept away a tangle of laborious calculations in invariant theory, including much of the life work of the mathematician Paul

Gordan. Gordan's initial reaction to Hilbert's proof has become famous ('that is not mathematics; that is theology'); but he later came to accept that 'theology also has its advantages'.

Hilbert himself believed that among all proofs of a given theorem there must always be one that is simpl*est*. He said so explicitly, in the statement of an unpublished problem, recently discovered in his *Nachlass*, that he intended to include with the twenty-three that make up the published text of his lecture to the Second International Congress of Mathematicians. As quoted in [Thiele, 2001, p. 16], that twenty-fourth problem 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'.

It is fair to say that if the criteria Hilbert had in mind were *formal* ones, as might be expected in the context of his proof theory, then little attention has so far been paid to the problem of proof simplicity. Within the restricted field of pure logic, some progress has been made by those concerned with automated theorem proving. (For an optimistic report on past achievements and future prospects in that area, see [Thiele and Wos, 2002].) Beyond logic itself, however, as discussed in the previous section, there has not even been consensus as to when two formal proofs with the same premises and conclusion represent conceptually *equivalent* arguments. Accordingly, judgments of whether one proof in ordinary mathematics is simpler than another, and in what sense, have up to now been based mainly on informal intuitive criteria such as those listed above.

Another important reason for giving alternative proofs is:

(3.3) *to demonstrate the power of different methodologies*,

for research, pedagogical, or ideological purposes.

The book [Fine and Rosenberger, 1997] is a fine pedagogical example. In it, no fewer than twelve different proofs of the Fundamental Theorem of Algebra are discussed, employing such disparate methodologies as advanced calculus, the theory of field extensions, Galois theory, complex analysis, and algebraic topology. The aim of the text is to introduce undergraduates to 'a great deal of non-elementary mathematics, all centered on a single topic'.

An example of an ideological text written to promote a particular research methodology, satisfying both aims (3.1a) and (3.3), is Errett Bishop's *Foundations of Constructive Analysis* [Bishop, 1967], subsequently revised as [Bishop and Bridges, 1985]). Described by its author as 'a piece of constructivist propaganda', that text demonstrated forcefully how large a part of abstract analysis can be developed within a constructive framework.

Both research and pedagogical purposes were served by Abraham Robinson's creation of non-standard analysis, which H. J. Keisler later

employed in his innovative calculus text [Keisler, 1986]. The historical fact that the theorems of calculus were first proved by infinitesimal methods bears witness that proofs employing infinitesimals are often more direct and perspicuous than those couched in terms of limits. It was the *concept* of infinitesimals that seemed vague, and Robinson's principal aim in developing non-standard analysis, in accord with goal (3.1) above, was to provide a rigorous definition of infinitesimals and carry out proofs based upon it in order

(3.4) *to provide a rational reconstruction (or justification) of historical practices*.

The development of a rigorous theory of generalized functions, in order to explain results that had been obtained by appeal to Dirac's delta 'function', is another example of (3.1a) and of (3.4).

One of the most important reasons for giving new proofs, already mentioned in connection with reason (3.2c), is

(3.5) *to extend a result, or to generalize it to other contexts*

in which the original proof fails to apply, by introducing new concepts or approaches.

Knorr's book [1975], cited earlier, provides an ancient paradigm for both (3.4) and (3.5). Its subject is the report in Plato's *Theatetus* that Theatetus's teacher, Theodorus, proved that if the area of a square is less than 17 and not a perfect square, then the side of that square is not commensurable with the side of a square of unit area. The question is: why the restriction to areas less than 17? The reason, Knorr conjectures, is that Theodorus's proof was *not* the modern one, based on the uniqueness of prime factorizations, but a simpler one based on whether a number is even or odd. That simpler proof (unlike the modern one) does not work for integers that are congruent to 1 mod 8, and so, apart from 9 (a perfect square), first fails for 17.

Other examples of (3.5) abound. Among them are Euler's proof of the infinitude of the primes—conceptually much more advanced than Euclid's, but applicable to a wider range of number-theoretic problems, such as proving the infinitude of primes in arithmetic sequences; Henkin's proof of Gödel's completeness theorem for first-order logic, which removed the restriction that the underlying language be countable and introduced a technique (the method of constants) that is fundamental in other model-theoretic constructions; and the many theories of integration, developed to refine theorems concerning the Riemann integral and to extend them to various contexts beyond integration over intervals on the real line.

The reasons enumerated above for giving new proofs are for the most part *practical* ones, aimed at resolving lingering doubts, reducing computational effort, exhibiting explicit solutions to problems, widening the scope of a theorem's applicability, or rendering a proof comprehensible

with less intellectual effort or background knowledge. But there are other motivations for seeking new proofs that are of a more personal or *aesthetic* nature. For mathematics is a creative human endeavor, akin in some respects to mountaineering: mathematicians often seek to solve problems for the same reason that mountaineers are driven to climb mountains—because they are there; and as in mountaineering, being first to the summit is not the only worthwhile goal. It is also exciting

(3.6) *to discover a new route*

to reach it.

Some new proofs are thus presented purely as such, because they are deemed to be particularly *novel*, *ingenious*, or *elegant*, or (like the free climber who disdains mechanical aids) because they attain the goal through restricted means (as, for example, Mascheroni's proof that the compass alone suffices for all straightedge-and-compass constructions). Such proofs frequently involve clever insights of quite limited generality. Nonetheless, they are often prized for their beauty, as the compilation [Aigner and Ziegler, 2000] attests.

But perhaps the most important aesthetic reason for devising new proofs is

(3.7) *concern for methodological purity*

(which also exemplifies the practical aim (3.2c)).

The quest for methodological purity in mathematics has been the subject of recent study by Andrew Arana,[11] who defines purity of method in mathematics as 'a balance between the conceptual resources that are used to formulate and comprehend a problem and the conceptual resources used to solve it'. Proofs, that is, are 'pure' to the extent that they do not invoke notions extraneous to the concepts involved in the statement to be proved.

There are numerous examples throughout mathematical history of efforts to 'purify' proofs. A well-known example from antiquity is Euclid's attempt in the *Elements* to avoid using superposition wherever possible. Proofs of the Fundamental Theorem of Algebra that aim to minimize the use of analytical or topological concepts also fall within this category, as do efforts in mathematical logic to give proofs of model-theoretic results (such as the Compactness Theorem) without resort to syntactic considerations.

Perhaps the most outstanding indicator of the high regard mathematicians have for purity of method was the award, mentioned earlier, of a

[11] An outline of a lecture by him, containing some illuminating historical and technical insights on that topic, is available online (http//:www-personal.ksu.edu/∼aarana/talks/PennStateLogicSeminar2005.pdf). See also [Arana and Detlefsen, 2005].

Fields Medal to Atle Selberg. Among Selberg's many contributions, the award citation noted in particular his 'elementary proof of the prime number theorem (with P. Erdös), with a generalization to [Dirichlet's theorem on] prime numbers in an . . . arithmetic progression'.[12] The proof of Erdös and Selberg used notions from real analysis—which, as Arana points out, did not compromise its purity, since the statement of the prime-number theorem involves the real-analytic concept of logarithm. But one may question (as Arana does) whether Selberg's proof of Dirichlet's theorem, which also used real-analytic notions, is truly pure, since the statement of that theorem ('If a and b are relatively prime, then there are infinitely many primes that are congruent to a mod b') involves only simple number-theoretic notions.

One might more properly speak of *degrees* of purity. But there are also instances in which different notions of purity *conflict*. One example, not cited by Arana, is Desargues's Theorem in the plane, which states that if two triangles in the same plane are situated so that the lines joining their corresponding vertices are concurrent, then the corresponding sides, 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, however, be reconciled, since it can be proven that any planar proof of the theorem must invoke the *metric* notion of similarity.

It is worth noting, as well, that model-theoretic proofs of consistency, in which basic notions are given alternative semantic interpretations, inherently violate purity of method. That is probably the reason, as Jeremy Gray has suggested, why spherical geometry was not recognized as a model for non-Euclidean geometry long before the work of Bolyai and Lobachevski: 'spherical geometry [could] exist alongside [belief that negating the parallel postulate led to inconsistency] because spherical geometry treats of curved lines on curved surfaces and not of straight lines' [Gray, 1989, p. 171].

Note, too, that proofs that violate purity of method may possess merits of their own. For example, consideration of the singularities of a power series in which the variable is allowed to take on complex values may clarify *why* the radius of convergence of the corresponding real-valued power series is what it is—an insight that could not be obtained without introducing the complex perspective. In such a case, the more general context has greater explanatory power.

[12] In the technical sense of not involving complex analysis. That the proof was not elementary in the sense of being easy to follow is evidenced by the publication twenty years later of a simplified version entitled 'A *motivated* account of an elementary proof of the prime number theorem' ([Levinson, 1969], my emphasis).

(3.8) Finally, the existence of multiple proofs of theorems serves an overarching purpose that is often overlooked, one that is analogous to the role of *confirmation* in the natural sciences. For just as agreement among the results of different experiments heightens credence in scientific hypotheses (and so also in the larger theories within which those hypotheses are framed), different proofs of theorems bolster confidence not only in the particular results that are proved, but in the overall structure and coherence of mathematics itself. To paraphrase a remark C. S. Peirce once made with regard to philosophy, trust in mathematical results is based rather on 'the multitude and variety' of the deductions that lead to them than on 'the conclusiveness of any one' of those deductions. Mathematical reasoning forms not 'a chain, which is no stronger than its weakest link, but a cable', whose fibers, though 'ever so slender', are 'numerous and intimately connected' ([Peirce, 1868], reprinted in [Peirce, 1982–, p. 213]). Or as Max Dehn put it, more succinctly, 'Most [mathematical] results are so involved in the general web of theorems, they can be reached in so many ways, that their incorrectness is simply unthinkable' [Dehn, 1928].

Appendix: Proofs of the Fundamental Theorem of Arithmetic

The Fundamental Theorem of Arithmetic (FTA) states that any two factorizations of an integer into primes must be identical except for the order of the factors. The FTA implies as a corollary the Basic Divisibility Lemma (BDL), which asserts that if a prime divides a product it must divide one of the factors. The BDL, restricted to products of two factors, was proved by Euclid as Proposition VII, 30 of the *Elements*. Related results are Proposition VII, 32 ('Any positive integer is either prime or is divisible by some prime.') and Proposition IX, 14 ('If a number be the least that is measured by [three distinct] prime numbers, it will not be measured by any other prime number except those originally measuring it.'). Euclid does not consider products of more than three primes, nor products involving repeated factors. Note, however, that repeated application of VII, 32 establishes the existence of a prime factorization for any positive integer, and Euclid's proof of IX, 14 can be applied to conclude that the representation of a number as a product of *distinct* primes is unique except for the order of the factors. To extend to products involving repeated prime factors one can then argue, as Gauss later did, that if a prime p appears to the power j in one factorization of a number n and to the power k in another, with $j \neq k$, then dividing by p^j will yield two factorizations of another number, at most one of which involves the factor p. Applying VII, 30 repeatedly then shows that p must in fact occur in neither (so $j = k$). Any other repeated prime factors may be similarly eliminated; so that only products of distinct prime factors need be considered.

Consequently, the BDL also implies the FTA, and we may distinguish proofs of the FTA that do *not* first prove the BDL from those that do. We may also note to what extent proofs of the FTA use induction, whether they invoke the concepts of least common multiple or greatest common divisor, and whether they employ the division algorithm, the Euclidean algorithm, or neither.

In proposition VII, 30 Euclid actually establishes a slightly stronger result than the BDL, namely, that if A divides BC and is relatively prime to B, then A divides C. The proof, however, is rather murky and involves the seemingly extraneous notion of ratios in lowest terms. A modern reconstruction that eliminates reference to ratios goes as follows: Let A be a prime that divides the product BC, say $BC = AD$. If A does not divide B, then A is prime to B, and C $\neq$ D. Consequently (VII, 21), A must be the least positive integer which, when multiplied by D, is a multiple of C. (That is, AD is the least common multiple of D and C.) For let F be the least such integer, and suppose $FD = EC$. By the division algorithm, $A = QF + R$ for some Q, R with $0 \leq R < F$. Hence $BC = AD = QFD + RD = QEC + RD$; so RD $= BC - QEC = (B - QE)C$. By the minimality of F, $R = 0$; so $A = QF$ and, since $C \neq 0$, $B = QE$. Thus Q is a common factor of A and B; so $Q = 1$, whence $A = F$. Then, again by the division algorithm, $C = qA + r$ for some q, r with $0 \leq r < A$. Thus $r = C - qA$; whence $rD = CD - qAD = CD - qBC = (D - qB)C$; that is, rD is a multiple of C. By the minimality property of A, $r = 0$; so A divides C.

This proof involves two applications of the division algorithm, and indirectly uses the notion of least common multiple. In contrast, a recent paper of Daniel Davis and Oved Shisha [Davis and Shisha, 1981] gives five variant proofs of Euclid's result that are remarkable for their economy of means. None of them makes any use of the division algorithm, and all are *reductios*, starting from the assumption that there is a triple (A, B, C) of positive integers for which both of the following properties hold:

$P_1(A, B, C)$: A divides BC but is relatively prime to B;

$P_2(A, B, C)$: A divides BC but does not divide C.

Using nothing more than subtraction[13] and the concepts involved in the theorem statement (divisibility and relative primality), it is very easy to see that for $i = 1$ or 2:

$$\text{if } P_i(A, B, C) \text{ and } B > A, \text{ then } P_i(A, B - A, C), \tag{1}$$

[13] It may be objected that the division algorithm itself is merely a matter of repeated subtraction. To implement that algorithm, however, requires a loop, in contrast to the unrepeated subtraction involved in the argument below. Moreover, as Davis and Shisha note, their proofs do not even involve the concept of remainder.

and

$$\text{if } P_i(A, B, C), \text{ say } BC = AD, \text{ then } P_i(B, A, D). \tag{2}$$

The five variants differ in what quantities involving the elements of the putatively existent triple are minimized: first A, then B; first B, then A; first C, then B; $A + B$; or $A + B + C$. In the last case, suppose (A, B, C) is a triple satisfying P_1 and P_2 for which $A + B + C$ is minimal. Then $A \neq 1$ (by P_2); so $A \neq B$ (by P_1). The minimality of $A + B + C$, together with (1), rules out $B > A$, while the minimality of $A + B + C$, together with (2), implies that $B > A$. Hence no such triple exists.

Two other proofs of the BDL are given in [Rademacher and Toeplitz, 1957, pp. 71–72], and in [Courant and Robbins, 1941, pp. 45–47]. The first of those begins by applying the division algorithm, once, to show that the least common multiple m of two numbers a and b divides every common multiple of them, and notes further that if $md = ab$, then d must divide both a and b. So if a prime p divides AB, let L be the least common multiple of p and A. Then since AB and pA are both common multiples of p and A, L divides AB—say $LE = AB$—and L divides pA— say $LF = pA$, where F is a common divisor of p and A. Since p is prime, either $F = p$ or $F = 1$. In the former case, p divides A. In the latter case, $LE = pAE = AB$; so $pE = B$, that is, p divides B.

The other proof invokes the Euclidean algorithm (*Elements* VII, 2, proved by *repeated* use of the division algorithm) to show that the greatest common divisor D of two positive integers, A and B, can be expressed in the form $mA + nB$ for some integers m and n (exactly one of which must be negative). If p is a prime that does not divide A, then the greatest common divisor of p and A is 1; so there are integers m and n for which $mp + nA = 1$. Hence if p divides AB, say $AB = pC$, then $mpB + nAB = mpB + npC = p(mB + nC) = B$; so p divides B.

Alternatively, as is common in texts on ring theory, the representation of D in the form $mA + nB$ may be established by first showing that the set I of *all* such linear combinations (where m and n range over all integers) is an ideal within the ring of integers. If D is then chosen to be an element of I for which $|D|$ is minimal, a single application of the division algorithm shows that D must divide every element of I (and in particular, must divide A and B). Since any common divisor of A and B must also divide every element of I (in particular, D), D must be a greatest common divisor (as must $-D$, whence D may be taken to be positive).

In contrast to the foregoing proofs of the BDL, inductive proofs by *reductio* of the FTA, by Ernst Zermelo and Gerhard Klappauf (originally published in [Zermelo, 1934] and [Klappauf, 1935]), are reproduced in

Arnold Scholz's text *Einführung in die Zahlentheorie*. (Zermelo's proof also appears on pp. 23–24 of [Courant and Robbins, 1941].)

Zermelo's stated aim was to show that even in elementary number theory it is quite possible to simplify the proofs, despite their elementary character. His proof runs as follows: By *reductio*, assume there are numbers with distinct prime factorizations, and let m be the least such, say $m = p_1 p_2 \dots p_k = q_1 q_2 \dots q_s$, where $p_1 \leq p_2 \leq \dots \leq p_k$ and $q_1 \leq q_2 \leq \dots \leq q_s$. By the minimality of m, $p_1 \neq q_1$, so without loss of generality we may suppose that $p_1 < q_1$. Then the number $n = m - p_1 q_2 \dots q_s = p_1(p_2 \dots p_k - q_2 \dots q_s) = (q_1 - p_1)(q_2 \dots q_s)$ is less than m, and so must possess a unique prime factorization. Since p_1 is less than every q_i, it must therefore divide $q_1 - p_1$. But then p_1 would divide q_1, which is prime. Since $1 < p_1 < q_1$, that is impossible.

Klappauf, in turn, published his proof to show that the method by which Zermelo produced the counterexample n to the minimality of m could be further simplified. Taking m as in Zermelo's proof, he considered the remainders r_i, $i = 1, \dots, s$, obtained by dividing each q_i by p_1 . We have $q_i = a_i p_1 + r_i$, where each $r_i < p_1$. Since p_1 is less than each q_i, each a_i is positive; and since each q_i is a prime different from p_1, each r_i is also positive. Hence $m = q_1 q_2 \dots q_s$ can be written as $m = Ap_1 + R$, where $R = r_1 r_2 \dots r_s$ and A and R are both positive. Since p_1 divides m, it must also divide R. But p_1 cannot divide any r_i; so factoring each r_i into primes yields a factorization of R that is distinct from the factorization involving p_1. But $R < m$; so that contradicts the minimality of m.

Neither Zermelo's nor Klappauf's proof invokes the BDL, and Zermelo's does not invoke the division algorithm either. But in Klappauf's proof, $r_1 = q_1 - a_1 p_1 \leq q_1 - p_1$, and $r_i < q_i$ for $i \geq 2$; so the number R used to contradict minimality there is less than the number $n = (q_1 - p_1)q_2 \dots q_s$ used for that purpose in Zermelo's proof.

Consideration of these various proofs illustrates not only their structural differences, but some of the motivations for giving alternative proofs (simplicity, minimization of conceptual prerequisites, presentation of new approaches, and generalization to broader contexts) discussed earlier in this paper. In particular, Gauss's proof extended Euclid's; the proof in Courant and Robbins's text, as well as those by Zermelo and Klappauf, and especially that of Davis and Shisha, exhibit various forms of simplification; Davis and Shisha's proof is an exemplar of methodological purity; and the proof given by Rademacher and Toeplitz, in line with aim (3.6), employs a conceptual idea (that of representing the greatest common divisor of two integers as an integral linear combination of them), distinct from those in the other proofs, which leads directly to the BDL and has proved fruitful in the broader context of ring theory.

REFERENCES

AIGNER, MARTIN, and GÜNTER M. ZIEGLER [2000]: *Proofs from the Book*. 2nd ed. Berlin: Springer.

ARANA, ANDREW, and MICHAEL DETLEFSEN [2005]: 'Purity of proof'. (Preprint)

BARWISE, JOHN [1989]: 'Mathematical proofs of computer system correctness', *Notices of the American Mathematical Society* **36**, 844–851.

BISHOP, ERRETT [1967]: *Foundations of Constructive Analysis*. New York: McGraw-Hill.

BISHOP, ERRETT, and DOUGLAS S. BRIDGES [1985]: *Constructive Analysis*. Berlin and New York: Springer.

BUSS, SAMUEL [1991]: 'The undecidability of k-provability', *Annals of Pure and Applied Logic* **53**, 75–102.

CARBONE, ALESSANDRA [1997]: Review of [Buss, 1991], *Journal of Symbolic Logic* **62**: 1480–1481.

COURANT, RICHARD, and HERBERT ROBBINS [1941]: *What is Mathematics?* London, New York and Toronto: Oxford University Press.

DAVIS, DANIEL, and OVED SHISHA [1981]: 'Simple proofs of the fundamental theorem of arithmetic', *Mathematics Magazine* **54**, 18.

DEHN, MAX [1928]: 'The mentality of the mathematician. A characterization', *The Mathematical Intelligencer* **5** (1983), No. 2, 26.

DOŠEN, KOSTA [2003]: 'Identity of proofs based on normalization and generality', *The Bulletin of Symbolic Logic* **9**, 477–503.

FEFERMAN, SOLOMON [1975]: Review of [Prawitz, 1971], *The Journal of Symbolic Logic* **40**, 232–234.

—— [1978]: 'The logic of mathematical discovery vs. the logical structure of mathematics', in *PSA 1978*, vol. 2. East Lansing, Michigan: Philosophy of Science Assocation, pp. 309–327. Reprinted in [Feferman, 1998], pp. 77–93.

—— [1998]: *In the Light of Logic*. New York and Oxford: Oxford University Press.

FINE, BENJAMIN, and GERHARD ROSENBERGER [1997]: *The Fundamental Theorem of Algebra*. New York: Springer.

GRAY, JEREMY [1989]: *Ideas of Space, Euclidean, Non-Euclidean and Relativistic*. 2nd ed. Oxford: Oxford University Press.

KEISLER, H. JEROME [1986]: *Elementary Calculus: An Infinitesimal Approach*. Boston: Prindle, Weber & Schmidt.

KLAPPAUF, GERHARD [1935]: 'Beweis des Fundamentalsatzes der Zahlentheorie', *Jahresbericht der Deutsche Mathematiker-Vereinigung* **45**, 130.

KNORR, WILBUR R. [1975]: *The Evolution of the Euclidean Elements*. Dordrecht and Boston: D. Reidel.

LEVINSON, NORMAN [1969]: 'A motivated account of an elementary proof of the prime number theorem', *American Mathematical Monthly* **76**, 225–244.

MANCOSU, PAOLO [1996]: *Philosophy of Mathematics and Mathematical Practice in the Seventeenth Century*. New York and Oxford: Oxford University Press.

—— [2001]: 'Mathematical explanation: problems and prospects', *Topoi* **20**, 97–117.

MANCOSU, PAOLO, KLAUS JØRGENSEN, and STIG PEDERSEN, eds. [2005]: *Visualization, Explanation and Reasoning Styles in Mathematics*. New York: Springer.

PEIRCE, CHARLES SANDERS [1868]: 'Some consequences of four incapacities' *Journal of Speculative Philosophy* **2**, 140–157.

—— [1982–]: *Writings of Charles S. Peirce: A Chronological Edition*. Edward C. Moore, ed. Bloomington, Indiana: Indiana University Press.

PÓLYA, GEORGE [1954]: *Mathematics and Plausible Reasoning* (2 vols.). London: Oxford University Press.

PRAWITZ, DAG [1971]: 'Ideas and results in proof theory', in J.E. Fenstad, ed., *Proceedings of the Second Scandinavian Logic Symposium*. Studies in Logic and the Foundations of Mathematics, Vol. 63. Amsterdam and London: North-Holland, pp. 235–307.

RADEMACHER, HANS, and OTTO TOEPLITZ [1957]: *The Enjoyment of Mathematics*. Princeton: Princeton University Press.

RAV, YEHUDA [1999]: 'Why do we prove theorems?' *Philosophia Mathematica*, (3) **7**, 5–41.

SHAPIRO, STEWART [1991]: *Foundations Without Foundationalism: A Case for Second-order Logic*. Oxford: Clarendon Press.

THIELE, RÜDIGER [2001]: 'Hilbert and his 24 problems', in Michael Kinyon, ed., *Proceedings of the Twenty-Sixth Annual Meeting of the Canadian Society for History and Philosophy of Mathematics*. Canadian Society for History and Philosophy of Mathematics, pp. 1–22.

THIELE, RÜDIGER, and LARRY WOS [2002]: 'Hilbert's twenty-fourth problem', *Journal of Automated Reasoning* **29**: 67–89.

ZERMELO, ERNST [1934]: 'Elementare Betrachtungen zur Theorie der Primzahlen', *Nachrichten von den Gesellschaft der Wissenschaften zu Göttingen (Neue Folge)* **1**, 43–46.