major theorems. The new ideas seem after a while to be simple and natural, and future generations build on them to arrive at mathematics that is deeper, more general and more unified. Yet there will always be concrete examples like curves, surfaces and numbers! They will never go out of style, and will always serve as a source of and testing ground for new directions and new ideas.

References

1. P. Deligne, La conjecture de Weil, I, Publ. Math. IHES, 43 (1974) 273–307.
2. B. Dwork, On the rationality of the zeta function of an algebraic variety, Amer. J. Math., 82 (1960) 631–648.
3. A. Grothendieck, Formule de Lefschetz et rationalite des fonctions L, Séminaire Bourbaki 279 (1965).
4. R. Hartshorne, Algebraic Geometry, Springer-Verlag, New York, 1977.
5. H. Hironaka, Resolution of singularities of an algebraic variety over a field of characteristic zero, Ann. of Math., 79 (1964) 109–326.
6. F. Hirzebruch, Topological Methods in Algebraic Geometry, Springer-Verlag, Heidelberg, 1966.
7. K. Kendig, Elementary Algebraic Geometry, Springer-Verlag, New York, 1977.
8. S. Lang and A. Weil, Number of points of varieties in finite fields, Amer. J. Math., 76 (1954) 819–827.
9. R. Walker, Algebraic Curves, Dover Publications, New York, 1962.
10. A. Weil, Foundations of Algebraic Geometry, 2nd ed., American Mathematical Society, Providence, RI, 1962.
11. _____, Sur les Courbes Algébriques et les Variétés qui s'en Déduisent, Hermann, Paris, 1948.
12. _____, Number of solutions of equations over finite fields, Bull. Amer. Math. Soc., 55 (1949) 497–508.

STRUCTURING MATHEMATICAL PROOFS

URI LERON
School of Education, University of Haifa, Haifa, Israel

1. Introduction. Mathematical proofs are normally presented in a step-by-step, "linear" fashion, proceeding unidirectionally from hypotheses to conclusion. While this age-old and venerable method may be well suited for securing the validity of proofs, it is nonetheless unsuitable for a second, highly important role of most presentations—that of mathematical *communication*.

In this article an alternative method, called the "structural method," is proposed. The method, triggered by recent ideas from computer science, is intended to increase the comprehensibility of mathematical presentations while retaining their rigor. The basic idea underlying the structural method is to arrange the proof in *levels*, proceeding from the top down; the levels themselves consist of short autonomous "modules," each embodying one major idea of the proof.

The top level gives in very general (but precise) terms the main line of the proof. The second level elaborates on the generalities of the top level, supplying proofs for unsubstantiated statements, details for general descriptions, specific constructions for objects whose existence has been merely asserted, and so on. If some such subprocedure is itself complicated, we may choose to give it in the second level only a "top-level description," pushing the details further down to

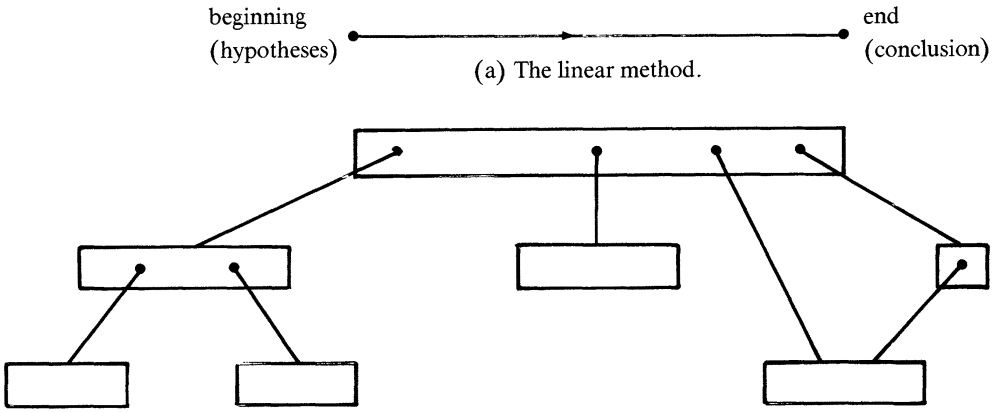
---

Uri Leron received his Ph.D. in 1972 from the Hebrew University of Jerusalem under S. A. Amitsur. In the years 1972–1980 he did research in ring theory and taught at the departments of mathematics of the University of Oregon, U.C.L.A., and the Technion—Israel Institute of Technology. Since 1980 he has been a Senior Lecturer at the School of Education, Haifa University. He is particularly interested in ways of communicating mathematical ideas that are not apparent from the mathematical formalism. He is also interested in the potential of computer programming for developing mathematical thinking. In connection with the latter he is currently carrying out an experimental project in which sixth graders learn programming through Turtle Graphics in the Logo language.

lower levels. And so we continue down the hierarchy of subprocedures, each supplying more details to plug in holes in higher levels, until we reach the bottom where (to borrow W. W. Sawyer's metaphor [**10**, p. 222]) all the leaks are plugged and the proof is watertight.

The top level is normally very short and free of technical (i.e., notational, computational, etc.) details. Thus it can be grasped at one glance, yielding an overview of the proof. (Note that the very term "overview" suggests view from the top.) The bottom level is quite detailed, resembling in this respect the standard linear proof. However, these details now appear only after their role in the proof is determined. Furthermore, they are now organized into conceptual units (the modules), each clearly and explicitly connected to its appropriate place in the total hierarchy. The intermediate levels facilitate a smooth transition from the generalities of the top level to the details of the bottom, from the global to the local perspective.

The two approaches are compared pictorially in Fig. 1. The linear method is represented by an oriented line segment (a), the structural method by a "structure diagram" (b). The structure diagram displays the levels, the modules and their interconnections. In each box, or module, the argument flows linearly, but it is very short and "flat" (no complex nesting patterns of sublevels); thus, again, it can be grasped at a glance. Incidentally, the description above applies not only to proofs, but to other mathematical procedures such as definitions, constructions, algorithms and examples as well.



(a) The linear method.

(b) The structural method.

FIG. 1. The two methods of presentation.

One may think of the structural approach as viewing the proof (which is at ground level) from a tall building. When viewing from the top we see the whole proof at a glance, but only in vague outline—no details can be discerned. As we descend the levels of the building, a zooming effect occurs: our view encompasses smaller and smaller segments of the proof, but these are seen with more and more clarity.

I anticipate a few objections some readers may have at this point. The first two objections, if carried to extremes, are, respectively, that what I am saying is wrong and that what I am saying is trivially true. More specifically, the first one may be expressed by the statement that "there is nothing wrong with the linear method; after all it was good enough for me and for generations of mathematicians." My answer is that we must look for insights in our students' behavior, not merely in ourselves. After all, most of us are successful survivors of the standard method, so we make up an extremely biased sample to draw conclusions from. (For a fierce attack on the linear method cf. [**6**, pp. 115–119].)

The second objection is represented by the following statement: "But your structural method is

what I've been doing in my classes all along." This statement, which has been expressed by some acknowledged excellent teachers, actually *supports* my argument. If the structural method provides a coherent and explicit system of presentation, whereby many of the "options" of good teachers become standard—it will have achieved its purpose. Note again that these options (e.g., a short overview of a long and complicated proof) not only become standard; they actually become part of the "official," formal proof.

A third argument that has been raised actually grants that knowledge of the structure of a proof is essential for its understanding. However (it goes on), this kind of knowledge is best left to the students to discover for themselves. My response to this subtler variant of the "it was good enough for me" argument is that indeed students should discover the structure for themselves; only they don't! And experience shows that this kind of task is beyond the capabilities of most undergraduates with standard mathematical training. In fact it may turn out to be an important teaching strategy within the structural method, to explicitly train students to structure linear proofs.

Finally, it should be stressed that the proposed modification bears directly only on one part of the teaching-learning process: the formal presentation of mathematical procedures as they commonly occur in lectures, textbooks and journal articles. By no means is it meant to replace the informal and artful devices that good expositors have always been employing to enhance the full and active participation of the student in this process, such as intuition, heuristics, personal metaphors, humor and even acting. After all, if the student is not listening, it matters little what we are saying!

2. Examples. I do not know of any way to prove (or disprove) my claims on the merits of the structural method. So in order to judge these claims we must rely on the powerful yet imprecise and subjective tools of our experience, intuition, reflections and observations; and a good way of activating these tools here is to carry out some mathematical "case studies."

On one level, these case studies are simply examples: they help us clarify and make more concrete the general concepts discussed previously. On a different level, they can function as a kind of mental experiment. That is, one can pretend to be a student reading these proofs for the first time (or an instructor teaching them to a class), and thus recapture mentally some of the real experience of learning (or teaching) by these methods.

Ideally we should study some long and complicated proofs, where the structure is buried deeply in the linear presentation (e.g., Brouwer's Fixed-Point Theorem), but we are limited here to theorems whose proofs are relatively short and elementary. Our examples will thus not be paradigmatic in the sense of exhibiting all the desired properties at once; rather, each will illuminate some of the issues only.

2.1. *The Infinite Number of Primes.* Euclid's original proof that there are infinitely many primes is so short and clear that it can be considered a one-level proof. So let us move further.

The odd prime numbers fall naturally into two classes: those leaving a remainder of 1 upon division by 4, and those leaving a remainder of 3. I shall call them "monadic" and "triadic" for short. Thus a number is monadic if it can be represented as $4k + 1$, and triadic if it can be represented as $4l + 3$ for some integers k and l.

It is natural to ask how the prime numbers are distributed between these two classes, and it turns out that there are infinitely many in *each* of them. (These are special cases of the following theorem of Dirichlet: *In any arithmetic progression* $\{an + b | n = 1, 2, 3, \dots\}$, *where a and b are relatively prime, there are infinitely many primes.*) We shall consider here only the simpler case, that of the triadic numbers.

THEOREM. *There exist infinitely many triadic primes* (*i.e., numbers of the form* $4k + 3$).

The first proof is patterned after a popular and well-known book.

Proof in the linear style. Consider a product of two monadic numbers:

$$(4k + 1)(4l + 1) = 4k \cdot 4l + 4k + 4l + 1 = 4(4kl + k + l) + 1,$$

which is again monadic. Similarly, the product of any number of monadic numbers is monadic.

Now assume the theorem is false, so there are only finitely many triadic primes, say $p_1, p_2, \ldots, p_n$. Define $M = 4p_1p_2 \cdots p_n - 1$.

If $p_i|M$, then $p_i|1$ since $p_i|4p_1p_2 \cdots p_n$. Since this is impossible, we conclude that no p_i divides M. Also 2 does not divide M as M is odd. Thus all M's prime factors are monadic, hence M itself must be monadic. But $M = 4p_1p_2 \cdots p_n - 1 = 4(p_1p_2 \cdots p_n - 1) + 3$ is clearly triadic—a contradiction. Thus the theorem is proved.

REMARK 1. Note that the student is led blindly and passively through the sequence of steps, which he or she must follow without ever being told the general plan of the proof and for what purpose the various steps are taken. Why, for example, should we "consider a product of two monadic numbers?" Why should we define a number M and why this particular way? Why should we check whether p_i and 2 divide M or not?

These questions should naturally arise in the mind of the alert student as he is going through the various steps of the proof, but the answers are not apparent in the presentation. And even the persistent student who does find the answers can do so only at the end of the proof. Worse yet, for many students, understanding (or learning) a proof has come to mean merely checking the validity of the deduction in each step. Thus they have stopped even asking these questions, let alone finding the answers.

REMARK 2. I hope you have read the proof with the student's eyes, not yours, since to a mature mathematician this proof is still very easy. However, similar frustration with the linear method can be experienced by mature mathematicians too, e.g., when reading a 15-page, strictly linear proof in a research journal.

Proof in the Structural Style.

Level 1. Suppose the theorem is false and let $p_1, p_2, \ldots, p_n$ be all the triadic primes. We construct (in Level 2) a number M having the following two properties:

(a) M as well as all its factors are different from $p_1, p_2, \ldots, p_n$;

(b) M has a triadic prime factor.

These two properties clearly produce a contradiction, as we get a triadic prime which is not one of $p_1, p_2, \ldots, p_n$. Thus the theorem is proved.

In the Elevator. (The elevator, as a metaphor for the intermediate process of descending in levels, offers a convenient place to discuss heuristic and other informal issues concerning the next level.)

How shall we approach the definition of M? In light of Euclid's classical proof, it is natural to try $M = p_1p_2 \cdots p_n + 1$. This indeed meets requirement (a), but not (b). In fact, since for all we know M itself may turn out to be prime, it must be triadic to meet (b).

Thus a natural second guess is $M = 4p_1p_2 \cdots p_n + 3$. However, this has another "bug": since one of the p_i's is 3, M is divisible by 3, in violation of (a). But this bug, once discovered, is easy to fix: simply eliminate 3 from the product in the bugged definition.

Level 2. Let $M = 4p_2 \cdots p_n + 3$ (we assume $p_1 = 3$). We show that M satisfies the two requirements from Level 1.

Requirement (a) means that no p_i should divide M. Indeed, $p_2, \ldots, p_n$ do not divide M as they leave a remainder of 3; and 3 does not divide M as it does not divide $4p_2 \cdots p_n$.

As for requirement (b), suppose on the contrary that all of M's prime factors were monadic. Then M, as a product of monadic numbers, would itself be monadic (Lemma, Level 3)—a contradiction. Thus (a) and (b) are satisfied.

Level 3: LEMMA. *A product of monadic numbers is again a monadic number.*

(The proof is as given above.)

REMARK 3. Level 1 demonstrates two important features of the top level. First, it can be grasped at one glance, and second, it gives the essence of the proof. Consider for example the introduction of M. What is stressed is its role in the proof, the properties it should enjoy, and how these are used to achieve the goal. The actual, detailed construction, as well as the proof that M satisfies the required properties, are pushed down to a lower level. Thus the top level indeed gives a global view of the proof.

REMARK 4. The "elevator" trick is a side benefit of the structural method, not an "official" part of it. Authors and instructors often like to inject informal remarks (mainly of a heuristic character) into the ongoing formal proof. In the linear method, where the argument is presented as an unbroken sequence of steps, there is no natural place to do it. So authors have been using various display tricks such as footnotes, different font, brackets and so on, to distinguish these remarks from the formal proof. In lectures, students are often confused as to what constitutes the formal proof and what are the auxiliary remarks.

In the structural approach, since the proof is divided anyhow into independent, relatively short modules, it is always possible to collect all the informal remarks and present them during the "elevator trip," that is, while descending to the next level.

REMARK 5. The successive debugging in the search for M illustrates an important point. I often ask my students to keep a protocol of the *process* of finding a proof, besides the final product. From these protocols it appears that no one reaches the correct M without going through some debugging, roughly of the kind given here. Since this is an essential part of the creative process of *doing* mathematics, I believe it should be made explicit and discussed with the students.

REMARK 6. Compare the location of the lemma on the product of monadic numbers in the two approaches. In the linear approach it appears in the beginning of the proof (or sometimes as a separate lemma *before* the main proof). The reader can have no idea why it is needed, but has to follow its proof nonetheless, and then keep the unused lemma in memory until it is needed at the *end* of the proof. In the structural approach the lemma is only announced when the need for it arises in the main proof. The lemma is used to complete the main proof and only then, in a lower level, its proof is given.

REMARK 7. The case of the monadic primes is more involved and will not be discussed here. However, it should be pointed out that in the structural method we have at least a clear beginning; in fact, the top level of the proof remains exactly the same, except for the exchange "triadic" ↔ "monadic." (One says that the two proofs exhibit a "top-level similarity.") The difficulty, as well as the "meat" of the proof, lie in the lower levels: the number M, promised in the top level, is not easy to deliver.

2.2. *A Theorem on Limits*

THEOREM. *If* $\lim_{x\to a} f(x) = L$ *and* $\lim_{x\to a} g(x) = M$, *then* $\lim_{x\to a} f(x)g(x) = LM$.

The following proof is not a caricature of the linear style; it is written strictly according to the "official" principles of that style and, in fact, is taken from a real calculus textbook.

Proof in the Linear Style. Let $\epsilon > 0$ be given, and let η be the smaller of $\sqrt{\epsilon/3}$ and $\epsilon/3(1 + |L| + |M|)$. Since $\lim_{x\to a} f(x) = L$, there exists a $\delta_1 > 0$ such that $|f(x) - L| < \eta$ whenever $0 < |x - a| < \delta_1$. Similarly, there exists a $\delta_2 > 0$ such that $|g(x) - M| < \eta$ whenever $0 < |x - a| < \delta_2$. Let δ be the smaller of δ_1 and δ_2. Now if $0 < |x - a| < \delta$, then $0 < |x - a| < \delta_i$, $i = 1, 2$, and so we have:

$$\begin{aligned}|f(x)g(x) - LM| &= |L(g(x) - M) + M(f(x) - L) + (f(x) - L)(g(x) - M)| \\ &\leqslant |L||g(x) - M| + |M||f(x) - L| + |f(x) - L||g(x) - M| \\ &< |L|\epsilon/3(1 + |L| + |M|) + |M|\epsilon/3(1 + |L| + |M|) + \sqrt{\epsilon/3} \cdot \sqrt{\epsilon/3} \\ &\leqslant \epsilon/3 + \epsilon/3 + \epsilon/3 = \epsilon. \qquad \text{q.e.d.}\end{aligned}$$

Fortunately, many instructors know better. They let the student in on the secret of how these mysterious quantities η and δ are actually *discovered*. But in so doing the direction of the argument is reversed, and eventually they have to abandon this unorthodox discussion and recast the official proof in more-or-less the form above (or at least mention that this recasting *should* be done).

The argument in the following structured proof resembles this informal discussion but at the same time it is quite formal and rigorous. Thus the structural approach brings closer the human process and the formal-deductive one.

A Structured Proof.

Level 1. Let $\epsilon > 0$ be given. We find (in level 2) a $\delta > 0$ such that $0 < |f(x)g(x) - LM| < \epsilon$ whenever $0 < |x - a| < \delta$. Thus the theorem is proved.

In the Elevator. We have to show that the expression $|f(x)g(x) - LM|$ can be made as small as we please. To this end, we try to bound it by an expression we *know* can be made small. Such expressions are $|f(x) - L|, |gx) - M|$ and multiples of these by a constant and by each other. After some trial and error the following expression emerges:

$$(*) \quad f(x)g(x) - LM = L(g(x) - M) + M(f(x) - L) + (f(x) - L)(g(x) - M).$$

Level 2. Using the equality $(*)$ we have

$$\begin{aligned}|f(x)g(x) - LM| &= |L(g(x) - M) + M(f(x) - L) + (f(x) - L)(g(x) - M)| \\ &\leqslant |L||g(x) - M| + |M||f(x) - L| + |f(x) - L||g(x) - M|.\end{aligned}$$

We find a $\delta > 0$ (in Level 3) such that whenever $0 < |x - a| < \delta$, each of the terms on the right-hand side is smaller than $\epsilon/3$. Thus the left-hand side is smaller than ϵ, as required.

In the Elevator. To get $|L||g(x) - M| < \epsilon/3$, we try to make $|g(x) - M| < \epsilon/3|L|$. However, there is a bug here: this only works if $L \neq 0$. One way of correcting this bug is to replace $|L|$ by $1 + |L|$. The case of $|M||f(x) - L|$ is similar. Finally, to get $|f(x) - L||g(x) - M| < \epsilon/3$, we make each of the factors smaller than $\sqrt{\epsilon/3}$.

Level 3. We choose positive $\delta_1, \delta_2, \delta_3, \delta_4$ such that the following hold:

$$\begin{aligned}&|f(x) - L| < \epsilon/3(1 + |M|) \quad \text{whenever} \quad 0 < |x - a| < \delta_1; \\ &|g(x) - M| < \epsilon/3(1 + |L|) \quad \text{whenever} \quad 0 < |x - a| < \delta_2; \\ &|f(x) - L| < \sqrt{\epsilon/3} \quad \text{whenever} \quad 0 < |x - a| < \delta_3; \\ &|g(x) - M| < \sqrt{\epsilon/3} \quad \text{whenever} \quad 0 < |x - a| < \delta_4.\end{aligned}$$

(Such δ_i's exist since L and M are the limits of $f(x)$ and $g(x)$ respectively.) Now let δ be the smallest of $\delta_1, \delta_2, \delta_3, \delta_4$, so that if $0 < |x - a| < \delta$ then $0 < |x - a| < \delta_i, i = 1, 2, 3, 4$. Then whenever $0 < |x - a| < \delta$, the expressions $|L||g(x) - M|, |M||f(x) - L|$ and $|f(x) - L||g(x) - M|$ all become smaller than $\epsilon/3$. Thus δ satisfies the requirements of Level 2.

REMARK. As seen from this example, structural proofs are longer to deliver, but (I believe) shorter to digest. In fact, they are longer because they contain more information (namely, the structure of the proof), and it is this very information which makes them more learnable,

illuminating and human. Thus in switching to structured proofs we simply agree to share with our students (or readers, etc.) more of what we know about the proof. And it is my belief that the loss in economy is more than balanced by the gain in learning.

2.3. *A High-School Level Problem...*

We next take up the following ruler-and-compass construction problem:

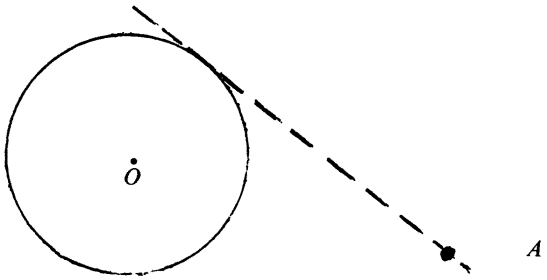


FIG. 2. A construction problem.

Construct a tangent to a given circle from a given point outside the circle.

The linear description of the construction is as follows (refer to Fig. 3).

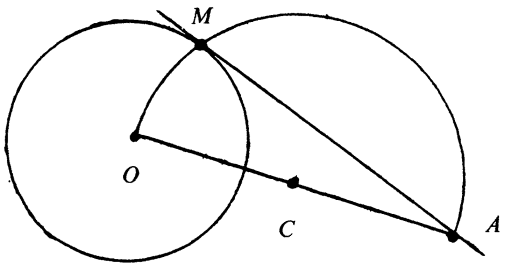


FIG. 3. The construction.

Draw the line segment OA and find its midpoint C. Around C as center draw a circle with radius $|OC|$. Let M be one point of intersection of the two circles. Then AM is the required tangent.

In general, the solution of construction problems in school usually proceeds in three stages. First, an analysis of the problem is carried out and the construction is discovered; second, the construction is described formally (in reverse order!); third, a proof is given that the construction indeed meets all the requirements.

In the structural presentation, all three stages are combined in one coherent process, thus again uniting the human (the first stage) with the mathematical (the second and third).

A Structured Proof.

Level 1. We find (in Level 2) a point M on the circle so that the line AM is perpendicular to the radius OM. Then (by a well-known theorem) AM is a tangent, as required.

In the Elevator. How shall we find the point M? We follow a well-known heuristic principle ([**9**, Ch. 1]): Look at the two properties determining M and construct the two corresponding loci; then construct M as the intersection of these two loci.

Level 2. Let M be a point of intersection of the given circle with the locus of points where the segment OA subtends a right angle. That is, M is a point of intersection of the given circle and the circle (constructed in Level 3) having OA as diameter.

Level 3. To construct the circle with segment OA as diameter, let C be the midpoint of segment OA, then draw the circle with C as center and $|OC|$ as radius. Clearly OA is a diameter in this circle, as required.

2.4. ... *And a More Advanced One*. The last example represented what seems, more or less, the lowest level of complexity in which arranging the proof in levels is still advantageous. While there is no corresponding upper limit (the more complex the proof, the more helpful is a structural presentation), we are nonetheless limited in our choice of case studies by the scope of this paper.

Pushing towards this limit, we now take up a theorem of central importance in linear algebra, namely the theorem on canonical forms for Hermitian operators and matrices. The theorem has several variants (the best known of which being, perhaps, the one on orthogonal diagonalization of symmetric matrices), and is a close relative of the Spectral Theorem. To save space, we state it directly in its more technical, ready-to-prove form.

THEOREM. *Let V be a finite-dimensional complex inner-product space, and let T be a Hermitian operator on V. Then V has an orthonormal basis of characteristic vectors of T.*

(Recall that T is Hermitian if $\langle Tu, v\rangle = \langle u, Tv\rangle$ for all vectors u and v in V, where $\langle x, y\rangle$ denotes the inner product of x and y.)

The following proof is self-contained modulo the standard material on operators and inner-product spaces; that is, it assumes no knowledge of Hermitian operators beyond the definition. See, e.g., [**2**], [**3**], or [**4**] for the standard facts, as well as for standard proofs of the theorem. Usually, the contents of this proof are spread out over several lemmas and theorems, thus achieving a measure of structuring even in the standard presentations. The structuring, however, is incomplete and the presentation proceeds *bottom-up*; that is, the auxiliary results (lower levels) appear first and the main argument (top level) last. (Note, however, the top-down commentary in [**3**, p. 301].)

A Structured Proof.

Level 1. Let U be the subspace of V spanned by all the characteristic vectors of T. We prove two assertions about U:

(a) $U = V$, i.e., the characteristic vectors of T span the whole space (Level 2.1);

(b) U has an orthonormal basis of characteristic vectors of T (Level 2.2).

Clearly, these two assertions yield the conclusion of the theorem.

REMARK. From here on the proof branches to two independent subproofs, rooted at 2.1 and 2.2. Note that while the presentation proceeds strictly top-down, it does not force a top-down *reading* of the proof. Some readers may prefer, for example, to read through the first branch (2.1) all the way to the bottom, then return to Level 2 and start on the second branch (2.2).

In what follows we have added a number (in parentheses) to the title of each module. This number is a back-reference to the "calling" module, i.e., the one where the present module is referenced. A more compact representation of the interconnections within the hierarchy is given in the structure diagram (FIG. 4).

Level 2.1 (1). To prove $U = V$ we prove the equivalent statement $U^{\perp} = \{0\}$, where $U^{\perp}$ is the orthogonal complement of U. This in turn will follow from the following two assertions:

(c) $U^{\perp}$ is T-invariant (Level 3.1).

(d) Every nonzero T-invariant subspace of V contains a characteristic vector of T (Level 3.2).

Since $U^{\perp}$ cannot contain a characteristic vector of T (this would contradict $U \cap U^{\perp} = \{0\}$), we must have $U^{\perp} = \{0\}$, hence $U = V$.

Level 2.2 (1). Let $\lambda_1, \ldots, \lambda_k$ be the distinct characteristic values of T, and let $U_1, \ldots, U_k$ be their respective characteristic spaces:

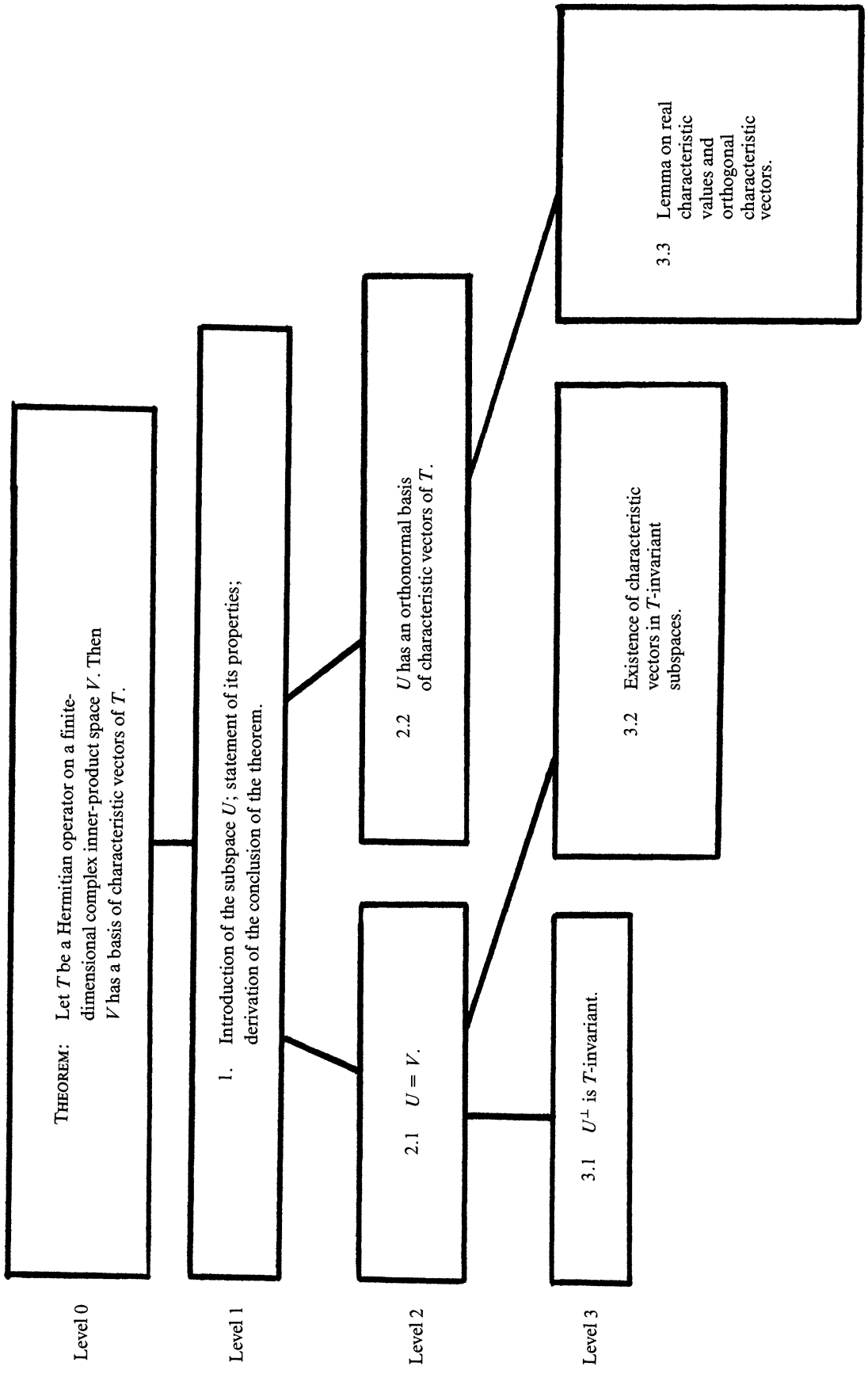


FIG. 4. Structure of the proof of the Theorem.

$$U_i = \{v \in V | Tv = \lambda_i v\}.$$

By the Gram-Schmidt orthogonalization process we can construct an orthonormal basis B_i to each subspace U_i, and it turns out that without any more care on our part, the set $B = B_1 \cup \cdots \cup B_k$ already forms an orthonormal basis for V. This is due to the following important lemma (proved in Level 3.3):

Characteristic vectors of T that belong to distinct characteristic values are orthogonal.

Thus if $i \neq j$, B_i and B_j are orthogonal, hence B is indeed an orthonormal basis for U. Since B consists of characteristic vectors of T, it satisfies all the requirements of Level 1 (b).

Level 3.1 (2.1). $U^{\perp}$ is T-invariant.

Level 3.2 (2.1). If W is a nonzero T-invariant subspace of V, then W contains a characteristic vector of T.

Level 3.3 (2.2). LEMMA. Let T be a Hermitian operator. Then

(a) *All the (complex) characteristic values of T are real.*

(b) *Characteristic vectors of T that belong to distinct characteristic values are orthogonal.*

We leave out the proofs of 3.1, 3.2 and 3.3, since Level 3 is the bottom level and the proofs appear in it in their standard (linear) form (see, e.g., [**4**, pp. 312–313]). Note, however, that it is only in 3.1 and 3.3 that the main hypothesis of the Theorem (namely that T is Hermitian) finally enters the proof. In contrast, 3.2 is true for operators on complex spaces in general.

REMARK. There is a similar theorem for normal operators, and the similarity is 3-level deep, counting the theorem itself as level 0 of the proof. Since 3.2 is true for operators on complex spaces in general, only 3.1 and 3.3 need to be adjusted.

We conclude this subsection with a structure diagram of the proof (Fig. 4), displaying the hierarchy of modules and their interconnections.

3. Summary: Benefits of the Structural Method.

3.1 *More Communicative Proofs*. The main benefit of presentations in the structural style is, hopefully, that ideas behind the formal proofs are better communicated. As we have seen, the main idea is given in the top level (or two), auxiliary ideas are packaged in autonomous modules, and the interconnections between the separate ideas are made explicit through the structure diagram.

By analyzing the foregoing examples, we can be a bit more specific about what is meant by the "main idea" of a proof, and the different treatment given to it by the two methods. The main idea often lies in the construction of a new, intermediate object, *the pivot*, to mediate between the hypotheses and the conclusion. (In our examples the pivots are the number M in 2.1, the numbers η, δ, δ_i in 2.2, the point M in 2.3, and the orthonormal basis of characteristic vectors, B, in 2.4.) Since the pivot occupies a central position in the proof and is directly connected to all its parts, it offers a vantage point from which to view the global architecture of the proof; and precisely this view is given in the top level of the structured proof (cf. the examples). Here the pivot is introduced by a statement of its essential properties and is immediately used to derive the conclusion of the theorem. The detailed definition, as well as proof of the postulated properties and questions of existence, are all pushed down to lower levels.

In the linear approach the pivot is treated poorly (from the learner's point of view) and its potential for revealing the architecture of the proof is wasted. Rather to the contrary, it is here where the proof most resembles pulling a rabbit from the hat (cf., e.g., the introduction of the number η in the linear proof of 2.2 or, for an extreme example, the element $f(a, b, c)$ in [**5**, p. 274]). The pivot is usually introduced near the beginning of the proof by a bare statement of its definition, which often appears extremely bizarre and complicated. Such definitions have an intimidating, even paralyzing, effect on many students when introduced too abruptly. The words

of Courant and Robbins, taken from a slightly different context ([**1**, p. 292]), are appropriate here: "There is an unfortunate, almost snobbish attitude on the part of some writers of textbooks, who present the reader with this definition without a thorough preparation, as though an explanation were beneath the dignity of a mathematician." Note, in contrast, the "untricking" effect a structured presentation has on the definition of the pivot: the actual detailed definition is given only after its role in the proof and the reasons for its particular form are made clear in the upper levels.

It should be remarked that the term "main idea" is used here more in the sense of an abstract or overview, and in no way is it meant to imply that mathematically it is the most important idea in the proof. For example, in the proof that there are infinitely many monadic primes (see Remark 7 of 2.1) the main *mathematical* idea may well be to take $M = (2p_1p_2 \cdots p_n)^2 + 1$, which is relatively a low-level idea. Still, I would maintain that awareness of the main idea in the above sense is important for an insightful learning, and is more likely to lead the student away from learning proofs and definitions by rote.

3.2. *Learning Activities.*

Several new, structure-related activities are available to the learner, or the instructor, in view of the structural method. Here is a sample:

—Given the higher levels of a proof, complete the lower levels.

—Take any proof from a standard textbook and find its structure (i.e., arrange the proof in levels). This is not easy, but it is highly valuable and rewarding; and it results in a much deeper understanding of the proof. (For some examples of this activity, see Section 2.)

—This is related to the previous one: Given a 10-page proof, describe it in one page.

—Given two theorems that show some similarity, determine how deep this similarity is; that is, to how many levels of the proof does the similarity extend, counting from the top.

All these activities can be used as self-learning activities, or they can be assigned as homework or test. Not only do they penetrate deeply into the particular subject matter, but they also encourage the learner to reflect on the process of theorem-proving itself.

An additional benefit of the structural method is that of *partial proofs*. The situation often arises, where presenting or studying a proof to the last detail is impossible or undesirable, due to insufficient level of mathematical preparation or motivation on the part of the audience, or simply due to time limitations. Consider, for example, the following cases: survey lectures or articles, mathematics courses for engineering students, an instructor scanning a pile of textbooks in preparation for a course, a mathematician scanning a pile of research articles. In all such cases the structural method allows one to conveniently choose a suitable level of detail for the particular situation, by simply ignoring (perhaps temporarily) some of the lower levels of the proof. This bears some similarity to Pólya's advice concerning the use of "incomplete proofs" ([**8**, "Why proofs?", pp. 219–221]), but note that our partial proofs can always be refined to a complete proof.

3.3 *Conclusion.* Iu. I. Manin has already referred to the human aspect of proofs when he said [**7**, p. 51]: "A good proof is one which makes us wiser." Clearly, then, a good *presentation* of a proof is one which makes the *listener* (or reader) wiser. It is my hope that presentations in the structural style have an improved chance of passing this difficult test.

Acknowledgment. I acknowledge gratefully the helpful comments from many colleagues on an earlier version of this paper.

References

1. R. Courant and H. Robbins, What is Mathematics?, Oxford University Press, New York, 1941.
2. P. R. Halmos, Finite-Dimensional Vector Spaces, 2nd ed., Van Nostrand, Princeton, 1958.
3. I. N. Herstein, Topics in Algebra, Blaisdell, Waltham, Massachusetts, 1964.

4. K. Hoffman and P. Kunze, Linear Algebra, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 1971.
5. N. Jacobson, Structure of Rings, revised ed., Amer. Math. Soc., Providence, RI, 1964.
6. M. Kline, Why the Professor Can't Teach, St. Martin's Press, New York, 1977.
7. Iu. I. Manin, A Course in Mathematical Logic, Springer-Verlag, 1977.
8. G. Pólya, How to Solve It, 2nd ed., Doubleday, Garden City, NY, 1957.
9. _____, Mathematical Discovery, vol. I, Wiley, New York, 1964.
10. W. W. Sawyer, A Concrete Approach to Abstract Algebra, Freeman, San Francisco, 1959.

WHO GAVE YOU THE EPSILON? CAUCHY AND THE ORIGINS OF RIGOROUS CALCULUS

JUDITH V. GRABINER
424 West 7th Street, Claremont, California 91711

Student: The car has a speed of 50 miles an hour. What does that mean?

Teacher: Given any $\varepsilon > 0$, there exists a δ such that if $|t_2 - t_1| < \delta$, then $\left|\frac{s_2 - s_1}{t_2 - t_1} - 50\right| < \varepsilon$.

Student: How in the world did anybody ever think of such an answer?

* * * * * * * * * * * * *

Perhaps this exchange will remind us that the rigorous basis for the calculus is not at all intuitive—in fact, quite the contrary. The calculus is a subject dealing with speeds and distances, with tangents and areas—not inequalities. When Newton and Leibniz invented the calculus in the late seventeenth century, they did not use delta-epsilon proofs. It took a hundred and fifty years to develop them. This means that it was probably very hard, and it is no wonder that a modern student finds the rigorous basis of the calculus difficult. How, then, did the calculus get a rigorous basis in terms of the algebra of inequalities?

Delta-epsilon proofs are first found in the works of Augustin-Louis Cauchy (1789–1867). This is not always recognized, since Cauchy gave a purely verbal definition of limit, which at first glance does not resemble modern definitions: "When the successively attributed values of the same variable indefinitely approach a fixed value, so that finally they differ from it by as little as desired, the last is called the *limit* of all the others" [**1**]. Cauchy also gave a purely verbal definition of the derivative of $f(x)$ as the limit, when it exists, of the quotient of differences $(f(x + h) - f(x))/h$ when h goes to zero, a statement much like those that had already been made by Newton, Leibniz, d'Alembert, Maclaurin, and Euler. But what is significant is that Cauchy translated such verbal statements into the precise language of inequalities when he needed them in his proofs. For instance, for the derivative [**2**]:

(1) Let δ, ε be two very small numbers; the first is chosen so that for all numerical [i.e., absolute] values of h less than δ, and for any value of x included [in the interval of definition], the ratio $(f(x + h) - f(x))/h$ will always be greater than $f'(x) - \varepsilon$ and less than $f'(x) + \varepsilon$.

Judith V. Grabiner has taught the history of science since 1972 at California State University, Dominguez Hills, where she is Professor of History. After getting a B. S. in Mathematics from the University of Chicago, she received her M. A. and Ph. D. (1966) at Harvard University. She is Book Review Editor of *Historia Mathematica*, Chairman of the Southern California Section of the Mathematical Association of America, and the author of *The Origins of Cauchy's Rigorous Calculus* (M. I. T. Press, 1981). In 1982–1983 she was Visiting Professor of History at U. C. L. A.

This paper is a revised version of a talk given at various mathematics colloquia, at the Summer Meeting of the Mathematical Association of America in Ann Arbor, Michigan, in 1980, and at the New York Academy of Sciences in March, 1982. Some of the research was supported by the National Science Foundation under Grant No. SOC 7907844.