Can Heuristic be Taught?

A review essay on *Infinitesimal calculus* by James M Henle and Eugene M. Kleinberg

PEGGY MARCHI

Any teacher of mathematics who has tried to teach his students techniques of solving problems cannot fail to have noticed that some students are much better problem solvers than others. Some students quickly learn how to go about solving problems or proving theorems (even if they do not always succeed in finding the solution or proof) and others seem unable to tackle problems where the solution cannot be found by rote, and unable to learn how to tackle such problems. Why are there these differences? Of course, some students are smarter than others and some students have more mathematical ability than others. But let us assume for the moment that we have a class of students all of whom are of above average intelligence and who have above average mathematical ability. Even in this group we will surely find that some students are better at solving problems than others. Some of the students seem to know effective problem solving rules and techniques — that I shall from now on call "heuristic" — and others don't. Why are there these differences?

Many mathematics teachers try to teach their students to be better problem solvers, believing that heuristic can be taught. Yet it is my observation, both as a teacher and a student, that a teacher is almost never able to improve a student's ability to solve problems. Despite the best efforts of the teacher — even one who consciously tries to improve the problem solving abilities of his students — at the end of the course the poor problem solvers are almost always still poor, even though they may have acquired some technical skills which they did not have at the beginning of the course. And the good problem solvers have probably not improved their problem solving abilities either.

Those who are concerned with teaching these students may seek an explanation of their failure in order that they may find a way to improve their teaching, or perhaps discover that what they are attempting is, in fact, impossible. Three hypotheses concerning, respectively, rules of heuristic, teachers of heuristic, and learners of heuristic, come immediately to mind. 1. Heuristic can be taught but the heuristic rules given in the text or imparted by the teacher are inadequate. Those students who are good problem solvers are using other heuristic rules in addition. 2. Heuristic can be taught, but even if the heuristic rules given in the text or imparted by the teacher are adequate, they are not taught in such a way that most students learn to apply them. 3. Heuristic cannot be taught. Good problem solvers may be born with the ability, or may teach it to themselves, but there is nothing we can do to enhance either.

The first section of this essay elaborates the difficulties one meets in trying to solve the problem of whether heuristic can be taught. The remaining sections discuss a new textbook which seems to teach heuristic — the philosophical difficulties raised in the first section notwithstanding.

I Can heuristic be taught? Any attempt to answer this question immediately suggests a prior question, namely, by what standard ought I to judge whether a student has been taught something, be it heuristic or anything else? By his ability to repeat what has been taught? By his ability to apply what he has been taught to familiar cases? By his ability to apply what he has been taught to unfamiliar cases? By his ability to criticize what he has been taught? If, unhappily, a student fails (by whatever standard) to learn, does that mean that the teacher has failed to teach? If a student succeeds (by whatever standard), has he succeeded because of the teacher, or independently of the teacher, or in spite of the teacher, or what?

Suppose that a student is taught (by whatever standard) heuristic, and suppose that a student nevertheless fails to solve an unfamiliar problem. Wherein lies the failure? The failure may be due to poor learning or poor teaching or both. But when the subject taught is heuristic, there are always other possible reasons for failure. For example, are the heuristic rules that we are teaching adequate? How can we tell if they are adequate? How can we tell, given a student's failure, whether the heuristic rules he applied are inadequate or whether the problem is just too hard for the student? This last is an especially difficult question to answer since heuristic rules never promise an automatic or guaranteed solution. So they can never be refuted by failure to solve a problem when using them because they never promise to give the power to solve all problems. What they *do* promise is however hard to articulate and consequently it is a hard problem to determine what counts as an adequate heuristic rule. Until we solve the problem of what counts as an adequate heuristic rule, we cannot tell whether failure to solve a problem while using a heuristic rule is a comment on the adequacy of the heuristic rule, or on the abilities of the problem solver, or on the impossibility of teaching heuristic, or some combination of all three.

So we seem caught in a vicious circle. We need to have adequate (by some standard) heuristic rules in order to determine whether or not these heuristic rules can be taught. But in trying to determine whether heuristic can be taught, we always come up against the possibility that failure is due to inadequate heuristic rules and not to the impossibility of teaching adequate ones. We can, of course, make a methodological decision to consider the rules adequate while we are testing whether they can be taught. This would break the vicious circle. But failure to teach heuristic is so common and the current state of heuristic so underdeveloped, that the cure of the vicious circle seems worse than the condition since our methodological assumption seems so obviously false. In other words, although the vicious circle seems to paralyze us and keep us from answering our original question, it seems at the same time better to admit how

little is known about the standards met by an adequate heuristic

So we find ourselves in the following situation: One way to determine if heuristic can be taught is to try and teach it and see if students become better problem solvers But for such a test to tell us whether heuristic can be taught, the heuristic rules which are taught must be adequate. That is, they must be rules which if followed will make people better problem solvers. How can we know if heuristic rules are adequate? Teach them to people and see if they become better problem solvers? Unfortunately, this move assumes what we are trying to test, namely that heuristic can be taught. Hence the vicious circle. We can break the vicious circle by fiat, i e. by declaring the heuristic rules adequate, at least for the purposes of testing But this move seems a poor one since we have good reason to suspect that our heuristic rules are not adequate. So we are stuck.

Suppose we try to break the vicious circle as follows: Declare as adequate those heuristic rules which *good* problem solvers follow. Then see if teaching these rules makes people better problem solvers Discovering the heuristic rules which good problem solvers use is the strategy which both Polya and Lakatos follow in their classic works on heuristic [1] This strategy has prompted fascinating historical studies of how mathematics grows and such historical studies strongly suggest that a greater knowledge of heuristic technique will improve our ability to solve problems. Yet as a strategy for determining whether heuristic can be taught, declaring as adequate the heuristic rules used by good problem solvers is of limited value. Admittedly, it has the advantage that it gives some reason other than methodological convenience for considering these heuristic rules adequate Yet the heuristic rules used by good problem solvers have been so little studied [2] that to declare these rules adequate and use the declaration to decide whether we can teach heuristic seems worse than just admitting our ignorance on both counts.

Given this situation, then, it seems that we can move if we are able to develop standards for what counts as an adequate heuristic which are independent of the question of whether we can teach an adequate heuristic. But the problem of standards for an adequate heuristic also turns out to be difficult. For example, although it was stated above that heuristic rules are not automatic and do not guarantee solutions to problems, this statement is more correctly a description of the heuristic rules we have rather than a known limitation on heuristic rules that may be discovered in the future. That is, we do not know whether we may, in fact, be able to find heuristic rules which are total algorithms for discovery It is an open question whether total algorithms for discovery could exist, although current opinion (in which I share) is against the possibility. Yet only a century ago people believed that a total discovery algorithm was possible and that it, in fact, existed in the shape of scientific induction.[3]

The problem is even more acute than this. Even if we found a completely successful algorithm for discovery, it would have to fit with our theories of whether we can have knowledge, and if so, how. For example, if we believe that all knowledge comes from scientific induction, then a non-inductive algorithm could not be a proper total algorithm, regardless of its success. If, on the other hand, we do not know whether anything can be known for sure, we have no way of judging whether an algorithm is a total algorithm or not There is an essential vagueness to the situation since our standards for a successful heuristic depend on our theories of knowledge, and our theories of knowledge are completely conjectural This fact resolves some of our difficulties while enhancing others It resolves the difficulty that we have no independent standard for a total algorithm while it makes more pressing the fact that even for partial algorithms we cannot escape from a certain looseness or vagueness [4]

If a heuristic rule needn't be a total algorithm, then how much better a problem solver must it make us to count as adequate? This question too leads to difficult problems. For, it seems that if we can specify how much better a problem solver a given heuristic rule will make us, or even that it will make at least one person a better problem solver at least once then it seems that our partial algorithm is in fact, a total algorithm under certain circumstances, contrary to our assumption. That is, setting any specific standard seems to demand that all adequate heuristic rules be total algorithms for discovery — at least under specific conditions. But if total algorithms for discovery are not possible (this is our current assumption) then this standard is unsatisfactory. Suppose then that we do not require that in order to be adequate a heuristic rule must make at least one person at one particular time a better problem solver Unfortunately this seems unsatisfactory, too, at least in the absence of any other standard, since we are interested in those methods and techniques which will in fact make people better problem solvers

What then should be the standards of adequacy for an heuristic rule? Since we have counter-intuitively (and perhaps mistakenly) barred the obvious standard, how then can we move? One way is to divorce our ability to solve problems from our ability to find solutions to problems. This sounds strange, even sophistic What is meant is that our knowledge of and intelligent use of heuristic techniques may be somewhat independent of our ability to find the solutions to the problems we need to solve. This is, of course, to some extent unsatisfactory since it begs the question of standards for heuristic; yet it — in my opinion rightly — reminds us that although luck favors the prepared mind, she can be capricious In other words, while we don't want to define the ability to solve problems purely as the ability to use known heuristic rules (since the known heuristic rules may not be adequate), at the same time we don't want to define it purely as success in finding solutions (since a person may be lucky without using heuristic)

Finally — and finally here means only that this is the extent of the author's knowledge of the difficulties and not necessarily the extent of the difficulties themselves — we may ask whether an adequate heuristic rule must contain an explanation (implicit or explicit) of how it improves our ability (by any standard) to solve problems, or is the rule adequate without such an explanation?

If all of this leaves the reader with spinning head, then the author has succeeded in her task of showing both the depth

and complexity of the difficulties which comprise the problem of teaching heuristic and at the same time the simpleness and intuitiveness of the idea — admittedly very difficult to articulate — that heuristic *can* be taught

II How then to approach these problems? I propose beginning in the middle by discussing the heuristic effectiveness of a new method for teaching the calculus, using infinitesimals, as presented in a new book, *Infinitesimal calculus* by James A. Henle and Eugene M. Kleinberg.[5]

We want to know whether heuristic can be taught and if so, how We are blocked in solving this problem because of the seeming mutual dependence of the teachability of heuristic and the adequacy of heuristic. The book I want to discuss claims to teach a better method for solving the theoretical problems of the calculus than its predecessors. That is, this new book claims to be heuristically better than its predecessors. This book may have found at least a partial solution to a problem which we could not solve, namely, how to teach heuristic, for it claims to succeed in teaching a better technique for solving problems in a specific area Therefore if we investigate this book's claims we may (1) discover the answer to our question; (2) develop our standards of what it means to teach heuristic; (3) decide whether better problem solving methods in a specific area improve our overall ability to solve problems; (4) decide whether the book lives up to its claims and use this decision to develop our standards for heuristic; or (5) some combination of the above. In other words, we can attack our problem from the middle and by doing so, develop our standards for teaching heuristic and our standards for determining the adequacy of heuristic and thus perhaps come somewhat closer to a solution to our problem

Infinitesimal calculus is an ambitious book. It seeks not only to teach the techniques of the calculus (e.g. differentiation, integration, etc.) and the important theorems and proofs of the calculus, but also to give students an appreciation both of the heuristic power of infinitesimals and of more general heuristic methods by showing how they are used to solve the theoretical problems of the calculus.

Infinitesimal calculus teaches the techniques of the calculus and the important theorems and proofs of the calculus by using a version of the method of infinitesimals. As is well known, infinitesimals — though simple and intuitive, and never giving a wrong answer in the calculus — were abandoned in the eighteenth century as a foundation of the calculus. They were abandoned because of the criticisms of Bishop Berkeley, which seemed to show their use to be based on an error in logic — essentially, changing the meaning of a term in the course of a deduction Berkeley argued that when using infinitesimals to compute, say, a derivative, one first assumes that the infinitesimals are non-zero increments given to the variables, and then that they are zero and hence can be taken away.

For example, the slope of the tangent to the curve $y = x^2$ at the point $x = 1$ would have been computed as follows: the ratio $[(1 + h^2) - 1^2]/h$ is an approximation to the true slope of the curve, where h is an infinitesimal, i e. not zero, but smaller than any real number. $[(1 + h)^2 - 1^2]/h$ can be reduced to $(2h + h^2)/h$, and then reduced still further to $2 + h$. Now since h is smaller than any real number, it can be considered to be zero and thus dropped. So the slope of the tangent to the curve $y = x^2$ at $x = 1$ is 2. Note that if h were zero all the way through the deduction we could not divide by it, but if it were non-zero throughout the deduction we could not drop it at the end of the deduction Thus we have not only changed the meaning of h during the course of the deduction (which is, of course, an error in logic) but doing so seems necessary in order to get the correct results The correct results come said Berkeley, from a compensation of errors. Berkeley's criticisms of the use of infinitesimals were universally accepted and other ways of developing the calculus were sought. In the mid-nineteenth century Weierstrass developed a finitistic technique which could serve as the basis of the calculus That is, he found a way to derive all the results of the calculus using only finite, non-zero quantities, whose values never need to change to zero in order to take away the increments added to the variables. This technique is called the epsilon-delta method. The epsilon-delta method is universally accepted as a correct basis for the calculus. The epsilon-delta method is, regrettably, complex and nonintuitive and lamentably difficult to teach.

For example, to compute the slope of the tangent to the curve $y = x^2$ at the point $x = 1$, we must first define the slope as the limit as h goes to zero of $[(1 + h)^2 - 1^2]/h$, then we must guess that this limit is 2, and finally we must show that 2 actually is the limit of $[(1 + h)^2 - 1^2]/h$ as h goes to zero. We do this by showing that for all real $\epsilon > 0$ there is a real $\delta > 0$ such that if $0 < |h| < \delta$ then $|[(1 + h)^2 - 1^2]/h - 2| < \epsilon$. *To show this we must guess* δ. In fact $\delta = \epsilon$ works. For, if $0 < |h| < \epsilon$ then $|[(1 + h)^2 - 1]/h - 2| = |2 + h - 2| = |h| < \epsilon$

Thus we have shown that we can make $[(1 + h)^2 - 1^2]/h$ as close to 2 as we like (i.e within ϵ of 2, but not equal to 2) by choosing the right bound for $|h|$ once ϵ is chosen Notice that in the epsilon-delta method, we must guess both the limit of the slope of the tangent and the right delta for a given epsilon, whereas in the proof using infinitesimals the limit was, so to speak, produced by the method of infinitesimals itself and no guessing was needed.

Thus compared to the proof using infinitesimals, the use of epsilons and deltas is both complex and unintuitive One source of the complexity and non-intuitiveness of epsilon-delta proofs is their indirectness. That is, although delta depends on epsilon, there is no algorithm for determining delta, given epsilon. Rather, the way in which delta depends on epsilon is *ad hoc* in each instance and must be discovered or guessed for each proof. The need to discover the way in which delta depends on epsilon makes epsilon-delta proof notoriously difficult to teach. Students usually do not understand why delta depends on epsilon at all, and if they understand that, they usually do not understand why delta must be guessed, i.e. why there is no formula for delta In epsilon-delta proofs we are expected to work on two levels at once — the concrete level of the mechanics of the proof (i.e finding delta) and the more abstract level where we specify the conditions which delta must meet (i.e. it must be a delta such that...) We are poorly prepared to combine these levels, perhaps because guessing delta with an end in

view seems neither quite rational (because guessing is not rational) nor quite fair (because if one is to guess, then only blind guessing is fair).

Teachers of calculus and textbooks of calculus, therefore, face the following problem: Infinitesimals are simple, intuitive, easy to teach, and their use never gives the wrong answer. Regrettably, their use also seems to require the use of fallacious reasoning, as Bishop Berkeley pointed out. The epsilon-delta method on the other hand is completely correct from the point of view of correct reasoning, but woefully difficult to teach. Teachers and textbooks of calculus are faced with a Hobson's choice of being either correct and rigorous but failing usually to teach the student to comprehend the reason for the correctness, or of being simple and intuitive but using a method which is suspect. Hardy's *Pure mathematics* is an example of the first; Raymond F. Coughlin's excellent *Applied calculus* is an example of the second.

Ideally, what we would like is a method of developing and teaching the calculus which is both simple and intuitive and at the same time consistent and rigorous. *Infinitesimal calculus* offers us this by re-introducing infinitesimals into mathematics in a way which does not commit errors in logic. It is based on the recent discoveries of Abraham Robinson in logic — in particular his method of non-standard analysis, which will be discussed at length later. *Infinitesimal calculus* succeeds in presenting the techniques, important theorems, and proofs of the calculus simply and intuitively, and at the same time, consistently and rigorously. *Infinitesimal calculus* is, in my opinion, much better than the usual calculus text.

In addition to teaching the techniques and theorems of the calculus, *Infinitesimal calculus* seeks to teach general heuristic methods by presenting students with problems in the theory of calculus. These are stated explicitly as questions, possible answers are discussed (usually very briefly in the main text and at greater length in the notes) and then infinitesimal analysis is used to develop the answers. Frequently, in the notes, alternative methods for answering the theoretical problems are given.

One example of its superiority in this respect is its discussion of uniform continuity. The authors explicitly state that we intuitively expect a sequence of continuous functions to converge to a continuous limit. They then show that this is not the case and raise the question: should we change our definition of continuity so that a sequence of continuous functions does converge to a continuous limit, or should we leave our definition of continuity alone and stop expecting a sequence of continuous functions to converge to a continuous limit? They explain that the first option involves Cauchy's stronger definition of uniform continuity (uniformly continuous functions converge to a continuous limit); the second, Lesbegue's weaker definition of measurable functions (continuous functions converge to a measurable limit).

Another aim of *Infinitesimal calculus* is to demonstrate the heuristic power of infinitesimals, that is, to show that infinitesimals are in fact simpler and more intuitive than epsilon-delta analysis. *Infinitesimal calculus* clearly succeeds in this aim. For example, to compute the slope of the tangent to the curve $y = x^2$ at $x = 1$ using the new method of infinitesimals, we let h equal an infinitesimal, i.e. a number which is not zero but which is smaller than any real number. Then we argue as we did in the proof which Berkeley criticized that the slope of the tangent will be $2 + h$. So far there is no error in logic. Now we say: We want the slope of the tangent to be a *real* number. But $2 + h$ is not a real number since h is infinitesimal. The real number which is closest to $2 + h$ is 2. So 2 is the slope of the tangent because 2 is the real number closest to $2 + h$. Notice that in this new method of infinitesimals, we do not let h become zero. Rather we change number systems: we compute the slope of the tangent by using infinitesimals but we interpret our answer in the real numbers. By operating first in one number system and then in another, we commit no error in logic.

This process of changing from one number system to another seems simple and intuitive. The ease with which we can now prove previously difficult results raises a puzzling question: *Why* are infinitesimals simpler and more intuitive than epsilon-deltas?

The simplicity of the method of infinitesimals derives from a very sophisticated use of symbolic logic. Explicitness concerning mathematics and silence concerning logic is the norm with mathematics texts and can be quite appropriate since questions of logic are usually concerned with the foundations of mathematics, i.e. with its justification. In a mathematics text, the business at hand is not the justification of the mathematics being taught but the mathematics itself. However, in *Infinitesimal calculus*, logic *is* part of the business at hand, since the heuristic superiority of infinitesimals is part of the business at hand. Thus, although *Infinitesimal calculus* is excellent as a calculus text, and excellent at showing the general method of growth in mathematics through the discussion of alternative solutions to theoretical problems, on the point where *Infinitesimal calculus* should be most interesting, original, and explicit, namely why infinitesimals are simpler and more intuitive than epsilon-deltas, it is regrettably silent. For so promising a book, this is a great disappointment.

The remainder of this review will be concerned, then, with questions concerning the heuristic superiority of infinitesimals. I shall first state briefly the sort of questions which need to be answered in order to explain why infinitesimals are heuristically superior. Then I shall discuss the points in this summary at greater length, by analysing a proof from *Infinitesimal calculus* as an illustration of my criticisms.

III Before beginning the critical part of this essay, I wish to put my criticisms into perspective. Reviewing *Infinitesimal calculus* has been a source of great intellectual satisfaction to me, for it is a book which speaks on many levels: mathematical, historical, heuristic, logical. It is also, obviously, a landmark in mathematics textbooks. On the mathematical and heuristic levels, it is excellent. And although, I have many complaints to make against it on the logical level, and a few on the historical level, in formulating my complaints I have learned an enormous amount both about logic and about the history of mathematics. In fact, the value of *Infinitesimal calculus* is perhaps best shown by

the fact that it merits lengthy and detailed criticism

The following are, briefly, my main criticism. (1) The concepts of metalanguage and object language are never explained, though their use in the book is crucial. (2) The question of whether the real numbers and the hyperreal numbers are structures, or languages, or both, is never adequately discussed, although again their appearance, sometimes as structures and sometimes as languages, is crucial to the book. (3) *Infinitesimal calculus* uses up to six different languages in a proof — or is it one metalanguage which contains two, or possibly four, object languages? In any case, the question of a mixture of languages is not discussed, although it is central to the book. (4) A knowledge of the properties of the reals and the ability to translate these properties into first order predicate calculus is assumed, but the question of how one might come to have such knowledge is not discussed.

The authors assume that the reader's intuition about logic will cope with these complexities, either by filling with his own reasoning the gaps in logic left by the authors, by jumping the gaps — i.e. going on with the book even though he cannot fill the gaps — or by not noticing the gaps and thus not being stopped by them. The authors provide a number of exercises designed to teach the reader how to operate in logic, and much help in working these exercises — i.e. heuristic designed to give the reader a feeling for logic. If the reader sees the gaps, his experience with the logical exercises will give him confidence that he could fill the gaps if he chose to, just as he is able to do with the mathematical exercises. If he does not notice the gaps, he can operate anyway since he has had the experience of doing the exercises. This seems to be the authors' attitude towards the reader's lack of knowledge of logic. It is an instance of the standard attitude of textbook writers towards gaps, especially gaps in the logical argument — that the reader will catch up.

Is such an attitude correct? Will the reader catch up? This is an empirical question and we will return to it.

Meanwhile, let us point out that an attitude of "The reader will catch up" begs many of the questions concerning the ability to solve problems, and the ability to teach people how to solve problems, that we raised in Part I. Since heuristic is not the main matter at hand in the standard mathematics textbook, begging such questions is permissible. But *Infinitesimal calculus* aims to develop the reader's intuition by showing the heuristic strength of infinitesimals, so heuristic is part of the business in hand and it does not seem right that the book should adopt a casual attitude at crucial points.

The authors also uncharacteristically miss the opportunity to explain how Abraham Robinson overcame a barrier to enable him to introduce the infinitesimals into the reals. The barrier, which we have referred to before, is that before Robinson no one knew how to use infinitesimals to get results in the calculus without committing logical errors. Robinson's method of avoiding the logical errors is hinted at by the authors, but never really explained. Suppose that we have a set of sentences, and suppose that we have two different ways of reading or interpreting these sentences, one using the real numbers and the other using the hyperreal numbers (i.e. the real numbers together with the infinitesimals). How do we make these interpretations? We interpret a sentence in first order predicate calculus by assigning a real number to each constant, assigning a property of the real numbers to each one place predicate, and assigning a relation between real numbers to each 2-, 3-, etc., place predicate. Since any sentence in first order predicate calculus contains only constants, predicates and relations, we can read it as a statement about real numbers by recalling the assignments we have made. In the same kind of way we can read a sentence as a statement about the hyperreal numbers.

We may think of the (uninterpreted) sentences as the language and the interpretations as structures which give the language meaning. But Robinson points out (taking the cue from Henkin) that we can see the interpretations as languages too. For the sentences interpreted as sentences describing real numbers can also be seen as a set of sentences whose individual constants are the names of real numbers, whose predicates are the names of properties of real numbers, and so on.

In *Infinitesimal calculus* first order predicate calculus interpreted in terms of real numbers is called the language L. L can then be thought of as a language describing all the properties of the real numbers which it can describe. As our authors point out, L is very limited and there are many properties of the real numbers which it cannot describe. For example, while it is true that "for all nonempty sets of (real) numbers B, if B has an upper bound, then B has a least upper bound", this fact about the real numbers cannot be expressed in first order predicate calculus, and hence cannot be described in L. ("We can quantify *numbers* in L, but not *sets of numbers*." page 22) In this book first order predicate calculus interpreted in terms of hyperreal numbers is called L^*.

Now we are ready to tackle the problem of how to introduce the infinitesimals into the reals without committing an error in logic. We recall that in the arguments that Berkeley criticized, an infinitesimal is first used as if its value is non-zero and then later as if its value is zero. Robinson noticed that a sentence like "The slope of the tangent is $2+h$" can be read in two different ways: as a statement about real numbers, where h is read as zero, and as a statement about hyperreals, where h is read as a non-zero infinitesimal. Both readings are valid, and no logical error has been committed.

And there is more. Since L and L^* are both languages, we can construct proofs in both L and L^*. And furthermore, we can prove sentences in L which can then be read in L^* and also sentences in L^* which can be read in L. Is it possible that if we prove a sentence in L^* that it is also proved in L? Robinson's answer is yes, so long as we put certain conditions on L^*. The crucial condition, according to Robinson, is that any sentence true in L must be true in L^*, and vice-versa.

When we began this explanation we started with a language and two interpretations. At that point we put no restrictions on the type of interpretations we might use, so we could have chosen two interpretations for which there were sentences true in one interpretation which were false in the

other. For example, the sentence "there is a y such that $y^2 = 2$" is true in the reals but false in the rationals. The condition which Robinson places on the interpretations — that every sentence true in L must also be true in L^* and vice-versa — enables us to prove things in one language and then transfer the result of the proof to the other language. So provided we arrange that every sentence in first order predicate calculus which is true for the reals is also true for the hyperreals, and vice-versa (and the hyperreals *can* be defined in such a way that this happens), then if we deduce, say, the slope of the tangent to $y = x^2$ at $x = 2$ in L^*, the result can be transferred to L and will give us a true statement about real numbers.

But have we constructed a proof? How do we know that a proof constructed in one language still holds when we change the interpretation? For if we read each line of the proof in the other interpretation, although each line will be true, its meaning may be such that it is not a logical consequence of the preceding lines and the sequence of statements cannot then be said to form a proof.

Our authors do not really explain this difficulty, although they give a hint of an explanation (page 31). The explanation rests on the fact that the only statements which are true in both L and L^* are those which do not name any of the hyperreals. That is, a sentence may be true in both languages if it contains *variables* which may be read as reals in L or as hyperreals in L^*, but if the sentence names a hyperreal it cannot be true in L since L contains no individual constants which correspond to hyperreals. The force of this fact will become clearer when we analyse a proof in the next section.

IV In this section I shall illustrate some of the points I have been making about *Infinitesimal calculus* by considering one of the proofs as an example. The theorem is that a continuous function on a closed interval is bounded. Our authors give the following proof:

PROOF: We would like the sentence

$$(*)\ \exists h \forall x(a \leq x \leq b \rightarrow -h < f(x) < h)$$

to be true in the real numbers, and our method simply will be to show it true in the hyperreals. It is easy, however, to see that $(*)$ is true in $\mathbb{HR}$ for if h is any infinite (positive) hyperreal then

$$\forall x(a \leq x \leq b \rightarrow -h < f(x) < h).$$

This is because given *any* x in $[a, b]$, $\boxed{x}$ must be in $[a, b]$, and so the continuity of f tells us that $f(x) \approx f(\boxed{x})$. However, $f(\boxed{x})$ is a real number (hence finite) and so $f(x)$ is finite. As h is infinite, $-h < f(x) < h$. So $\exists h \forall x(a \leq x \leq b \rightarrow -h < f(x) < h)$ *is* true in the hyperreals (taking h to be any positive infinite)— it must thus also be true in the reals. □

(*Note.* In the above proof, we assume that infinitesimals are smaller than all positive real numbers or larger than all negative real numbers, but not zero, and that infinite numbers, such as h in the proof, are larger than any positive real number or smaller than any negative real number. Both infinitesimals and infinite numbers are in the hyperreal number system.

This proof, like all the others in the book, is compact and simple. It is a remarkable fact that even knowing almost nothing about the hyperreals, one can at the very least follow the proof the first time through. This in itself is a great feat of clarity on the part of the authors. Yet there is more. The theorem which I have chosen for my example is moderately hard to prove using traditional epsilon-delta methods, yet using infinitesimals it seems easy. This ease in one system and difficulty in another raises a question which the authors do not answer. Given that the language L of the reals and the language L^* of the hyperreals are formally the same, why is the proof easy in L^* but hard in L?

But this formulation of the question can be dismissed by our authors, and rightly, for the proof is not in L^* in *Infinitesimal calculus* nor is it in L in traditional calculus books. We can improve the question by introducing the concepts of metalanguage and object language. An object language is a primary or basic language. (L and L^* are object languages in our discussion here.) A metalanguage is a language in which one can talk about an object language (English may be either an object language or a metalanguage, and so may other logical languages.) Our authors do not introduce these concepts, perhaps because they feel that students will switch naturally between object and metalanguage, using each when appropriate, and that to explain something which will be done naturally is superfluous. Yet not to make this distinction seems to obscure the possibility of explaining the heuristic superiority of Robinson's method of introducing infinitesimals into the reals.

Although the authors do not discuss the difference between object language and metalanguage, they do hint at the difference. They point out that L (the first order predicate calculus with individual constants "r" for each real number r) is limited and neither sufficient nor appropriate for more general use. They ask readers to notice that the book is not written in L (page 25). We may ask, then, in what language is the book written? In what language is the proof just quoted written?

One is tempted to say: English, of course. But let us be simple-minded for a moment here. If the proof is in English, then it would seem to follow that (*) is a sentence in English. Yet it is not. Although surrounded by a sentence in English, (*) is a sentence in L. Thus we may say, perhaps, in answer to the question about the language of the proof, that it is a metalanguage (in this case English) expanded to permit mention of sentences in L.

But the sentence in the seventh line of the proof is not an English sentence nor does it seem to be in L since it refers to infinite positive hyperreals and L cannot say anything about hyperreals. It seems, then, that it is in L^*. In fact, paradoxically, it is in both L and L^* (so is (*)), and we will return to this point shortly. Meanwhile we shall note here that the language of the proof

—is a metalanguage which allows mention of sentences of L and L^* (i.e. first order predicate calculus with individual variables h which refer to real numbers in L and hyperreal numbers in L^*)

—in which we can express facts about the mathematical structure of the real numbers and the mathematical structure of the hyperreal numbers
—in which we can describe the properties of the mathematical structure of the reals and mention them in L
—in which we can describe the properties of the mathematical structure of the hyperreals and mention them in L^*
—consider which sentences are true in both L and L^*
—choose whether we wish any of these sentences to be considered as belonging to L or to L^*

Now that we have described the metalanguage in which the proof is written, we can improve our question about the reason why Robinson's method is easier than the traditional one. Why is it easier to prove (*) by considering whether it is true of the hyperreals than by considering directly whether it is true of the reals?

I do not myself know the answer to this question, yet the following fact about the proof quoted above (which is true of all the proofs in *Infinitesimal calculus*) seems significant: Although we consider (*) as a sentence in L^*, the proof that (*) is true seems to take place in the metalanguage, and not in L^*. If the proof does, in fact, take place in the metalanguage and not in L^*, is this only a convenience or is it a restriction on L^*? If it is a restriction on L^*, why then do we need L^* at all for the proof? Why would we not use English only and forget about the logical complexities of L and L^*? I ask these questions not because I think that we can dispense with L and L^* but rather to point out that our authors do not explain why they are needed.

Perhaps such an explanation would be thought by the authors to be poor heuristic, i.e. perhaps they believe that there should be a limit to the complexity of the explanations given to beginning students of calculus in order not to overwhelm them from the start. Yet for at least one advanced beginner — myself — this lack of explanation made the book hard to understand at times and hard to follow. Even for real beginners, more explanation might have been better heuristic in a book which is otherwise so scrupulous in these matters.

An easy argument seems to show that the proof takes place in the metalanguage and cannot be translated into L^*. For suppose it could, then we would be able to express in L^* the (true) fact that there is a number which is not real (an infinitesimal, for example). But by definition of the hyperreals, all sentences true in L^* are also true in L. But when interpreted as a statement about real numbers, "There is a number which is not real" is false. So it seems that we cannot express in L^* the fact that a number is a hyperreal; or in L for that matter.

So the use of the metalanguage is essential for proofs using Robinson's method. More precisely, the inferences required for proofs can *only* be made in the metalanguage and not in the object language. Yet the authors of this book talk about whether a sentence is true in the hyperreals and express the sentence they would like to prove in L^*. But if the sentence cannot be proved in L^*, what is gained? Again, I do not know the answer. However, it seems significant that the authors alternate between treating the reals and hyperreals as languages and as structures.

The authors first present mathematical structures — the real numbers and some simpler ones — and the languages with which some of the properties of the structures can be described as if they were separate entities. Then (on page 22) they equate "structure" with "context"; for, as they say, "the truth of a statement depends on the context", i.e. on the structure being considered. So far so good. Then on page 23 they say, "Our languages themselves can be formed into structures." By "structure" here they do not seem to mean the intuitive concept of structure they have been using to that point, but a technical concept, in which a structure is not an entity but a language. This shift of meaning is not explained, and is presented as if it were no shift at all.

What is the nature of "context"? We understand a word or phrase or sentence when we know its context. For example, we understand "Time flies like an arrow" when we know the situation to which it refers. Sometimes this context is given by other sentences which may appear before, and sometimes after, the statement in question. These sentences are also sometimes called the "context". Is it the sentences, or the meanings of the sentences, which form the context? Our authors start by distinguishing the two possibilities and then identify them through the use of the term "structure".

Suppose a structure (in the intuitive sense) has a property which cannot be described in a particular language. This property is then not a part of the structure in the technical sense. Can this technical sense of structure lead us to a new intuitive idea of the entity we have been considering? On the one hand the answer seems to be yes, since we may now be able to imagine the structure without the unexpressible property. On the other hand the answer seems to be no, because the putative new structure comes only from paucity of language and not from a changed entity.

In what sense can the real numbers have a property which is not expressible in L? One example our authors cite is the completeness axiom, i.e. that every non-empty bounded set of real numbers has a least upper bound. And they point out (rightly) that the property cannot be expressed in L. But then what? Are we meant to think of the reals in our old intuitive way (e.g. with the axioms of completeness) but use only L in our development of the calculus? Or are we supposed to imagine a new structure (in the technical sense) also called the reals but having only the properties expressible in L? If so, is there more than one such structure for L? It seems that there must be, given the manner in which the hyperreals are constructed, yet our authors do not comment on this.

To conclude this discussion of questions which our authors do not answer, I shall raise again the question of why proof in the hyperreals is easy and proof in the reals hard in another form. Why is the intuitive picture of the hyperreals easy and the rigorous picture of the hyperreals difficult? In Chapter 4 our authors develop our intuitions of the hyperreals using the following intuitive definitions of infinitesimal and infinite number: *The absolute value of an infinitesimal is smaller than any real number but not zero; the absolute value of an infinite number is greater than any real number.*

With these intuitive working definitions, our authors show us that, for example, each real number "is surrounded by a cloud of non-standard numbers 'infinitely close' to it". (A non-standard number might be $r + h$ where r is a real number and h is infinitesimal.) In addition to the infinitesimals the hyperreals contain infinite numbers, which are the reciprocals of the infinitesimals. So the hyperreal line is both denser and "longer" than the real line. Hyperreal numbers have the calculational properties of the real numbers i.e. they can be added, subtracted, and so on, just like the real numbers.

Chapter 4 is easy to follow and intuitively appealing. By contrast, in Chapter 3, which according to our authors is skippable, we are presented with a different and much less intuitive picture of the hyperreals. Here the hyperreals are presented, *à la* Weierstrass, as sequences of real numbers. For example, a number r in the reals is the sequence $r,r,r,\ldots$ in the hyperreals. The sequence 1, ½, ⅓, ¼ ... is an infinitesimal hyperreal because it is not zero, i.e. not 0,0,0, ..., but at the same time is smaller than any positive real number, i.e. smaller than $r,r,r,\ldots$. Likewise, 1,2,3,4, ... is an infinite hyperreal because it is larger than any real number.

In Chapter 4 and onward all sentences true in L are true in L^*, so the properties of the reals expressible in L are also properties of the hyperreals and vice-versa. But in Chapter 3, the authors prove some of the properties of the hyperreals directly, that is, by considering the properties of sequences of real numbers. To do this, they introduce the idea of "quasi-big" sets, a species of infinite set. Two hyperreals, j and k, are equal if and only if the set of all n such that the nth terms of the sequences j and k ($j(n)$ and $k(n)$) are equal is quasi-big. Any relation R is true of some hyperreal j if and only if the set of all n for which $R(j(n))$ is true is quasi-big. For example, to prove that $j + k = k + j$ where j and k are hyperreals, we define addition of hyperreals as addition term by term of the sequences j and k. Since $j(n) + k(n) = k(n) + j(n)$ is true for all n, since the reals are commutative, the set of n is quasi-big, and so $j + k = k + j$.

At the end of Chapter 3 our authors prove that a sentence in L is true in the reals if and only if it is true in the hyperreals. Using this result we can recast our proof of commutativity in the hyperreals. Now we can argue that if j and k are hyperreals, $j + k = k + j$ is true for the hyperreals because (1) it is a sentence in L (as well as L^*) and (2) it is true in the reals, so (3) it is true in the hyperreals. The first proof of commutativity was not really hard, but this second proof is, by contrast, very easy.

Another example. Suppose we wish to prove that for hyperreals j and k, $j < k \rightarrow j \neq k$. (This example is an exercise from *Infinitesimal calculus* (page 29).) We can prove this directly for the hyperreals as follows: Assume that the set A of all n such that $j(n) < k(n)$ is quasi-big. We wish to show that the set B of all n such that $j(n) \neq k(n)$ is quasi-big. If we can show that A is a subset of B, then B is quasi-big, by the properties of quasi-big-ness. To show that A is a subset of B, if n belongs to A, then $j(n) < k(n)$; but $j(n) \neq k(n)$ since $j(n)$ and $k(n)$ are real numbers. So n belongs to B. So A is a subset of B. Q.E.D.

If, instead of proving this result about the hyperreals directly, we assume that a sentence in L which is true in the reals if and only if it is true in the hyperreals, then since $j<k \rightarrow j \neq k$ is true in the reals, it is true in the hyperreals.

The authors of *Infinitesimal calculus* stress that calculus proofs are easy in the hyperreals but hard in the reals. This point I concede completely. What I ask the authors of *Infinitesimal calculus* to explain is why this is the case.

Postscript

Which of the three methods of teaching calculus — the old, intuitive, infinitesimal method, the epsilon-delta method, or the new infinitesimal method — is the best for teaching calculus to beginners? In my opinion, the best method is the mathematically wrong method, i.e. the traditional infinitesimal method. The epsilon-delta method is no good for beginners because not only is it technically, i.e. mathematically, very difficult, but beginning students rarely understand the inadequacies in the foundations of the calculus that make the technical difficulties necessary. The new infinitesimal method is no good because it is logically very sophisticated, and hence difficult — even though mathematically it is very easy — and again the beginning student usually does not understand why the sophistication is necessary. The solution to this pedagogical problem, in my view, is to first teach calculus by the old infinitesimal method, then criticize the method *in detail* (something not done in any modern calculus book, regrettably including *Infinitesimal calculus*), and finally present an improved theory of the foundations of the calculus. Whether the epsilon-delta method or the new infinitesimal method should be the improved version is an open question. I incline towards the new infinitesimal method, especially since there is an excellent textbook, viz *Infinitesimal calculus*, to use in teaching it. Bertrand Russell complained that he was taught the old infinitesimal method instead of the epsilon-delta method, knowing that the old method was wrong. I think he might not have complained if his teachers had been Lakatosians who could have taught him that the wrong method was wrong.

References

[1] Georg Polya, *How to solve it*. 3rd Edition. Princeton, N.J.: Princeton University Press,
Georg Polya, *Mathematics and plausible reasoning*. 2 vols. Princeton, N.J.: Princeton University Press, 1954
Imre Lakatos, *Proofs and refutations*. London: Cambridge University Press, 1976

[2] In addition to the above:
Tobias Dantzig, *Number. the language of science*. New York: Macmillan, 1939
Arpád Szabó, Anfänge der Griechischen Mathematik. München; Oldenbourg, 1969
Otto Toeplitz, *Calculus. a genetic approach*. Chicago: University of Chicago Press, 1963

[3] Joseph Agassi, "Sir John Herschel's philosophy of success". In: Russell McCormach, *Historical studies in the physical sciences*. 1, 1-36 (1967)

[4] Joseph Agassi, "Sensationalism". In: *Science in flux*. *Boston Studies in the Philosophy of Science*. XXVIII, 93-119

[5] Cambridge, Mass: MIT Press, 1979