Purposes and Methods of Research in Mathematics Education

Alan H. Schoenfeld

Bertrand Russell has defined mathematics as the science in which we never know what we are talking about or whether what we are saying is true. Mathematics has been shown to apply widely in many other scientific fields. Hence, most other scientists do not know what they are talking about or whether what they are saying is true.

—Joel Cohen, "On the nature of mathematical proofs"

There are no proofs in mathematics education.

—Henry Pollak

The first quotation above is humorous; the second serious. Both, however, serve to highlight some of the major differences between mathematics and mathematics education—differences that must be understood if one is to understand the nature of methods and results in mathematics education.

The Cohen quotation does point to some serious aspects of mathematics. In describing various geometries, for example, we start with undefined terms. Then, following the rules of logic, we prove that if certain things are true, other results must follow. On the one hand, the terms are undefined; i.e., "we never know what we are talking about." On the other hand, the results are definitive. As Gertrude Stein might have said, a proof is a proof is a proof.

Other disciplines work in other ways. Pollak's statement was not meant as a dismissal of mathematics education, but as a pointer to the fact that the nature of evidence and argument in mathematics education is quite unlike the nature of evidence and argument in mathematics. Indeed, the kinds of questions one can ask (and expect to be able to answer) in educational research are not the kinds of questions that mathematicians might expect. Beyond that, mathematicians and education researchers tend to have different views of the purposes and goals of research in mathematics education.

This article begins with an attempt to lay out some of the relevant perspectives and to provide background regarding the nature of inquiry within mathematics education. Among the questions explored are the following: Just what is the enterprise? That is, what are the purposes of research in mathematics education? What do theories and models look like in education as opposed to those in mathematics and the physical sciences? What kinds of questions can educational research answer? Given such questions, what constitute reasonable answers? What kinds of evidence are appropriate to back up educational claims? What kinds of methods can generate such evidence? What standards might one have for judging claims, models, and theories? As will be seen, there are significant differences between mathematics and education with regard to all of these questions.

Purposes

Research in mathematics education has two main purposes, one pure and one applied:

- Pure (Basic Science): To understand the nature of mathematical thinking, teaching, and learning;
- Applied (Engineering): To use such understandings to improve mathematics instruction.

These are deeply intertwined, with the first at least as important as the second. The reason is simple: without a deep understanding of thinking,

Alan H. Schoenfeld is Elizabeth and Edward Conner Professor of Education at the University of California, Berkeley. His e-mail address is alans@socrates.berkeley.edu.

teaching, and learning, no sustained progress on the "applied front" is possible. A useful analogy is the relationship between medical research and practice. There is a wide range of medical research. Some is done urgently, with potential applications in the immediate future. Some is done with the goal of understanding basic physiological mechanisms. Over the long run the two kinds of work live in synergy. This is because basic knowledge is of intrinsic interest *and* because it establishes and strengthens the foundations upon which applied work is based.

These dual purposes must be understood. They contrast rather strongly with the single purpose of research in mathematics education, as seen from the perspective of many mathematicians:

- "Tell me what works in the classroom."

Saying this does not imply that mathematicians are not interested at some abstract level in basic research in mathematics education, but that their primary expectation is usefulness in rather direct and practical terms. Of course, the educational community must provide useful results—indeed, usefulness motivates the vast majority of educational work—but it is a mistake to think that direct applications (curriculum development, "proof" that instructional treatments work, etc.) are the primary business of research in mathematics education.

On Questions

A major issue that needs to be addressed when thinking about what mathematics education can offer is, What kinds of questions can research in mathematics education answer?

Simply put, the most typical educational questions asked by mathematicians—"What works?" and "Which approach is better?"—tend to be unanswerable in principle. The reason is that what a person will think works will depend on what that person *values*. Before one tries to decide whether some instructional approach is successful, one has to address questions such as: Just what do you want to achieve? What understandings, for what students, under what conditions, with what constraints? Consider the following examples.

One question asked with some frequency by faculty and administrators is, "Are large classes as good as small classes?" I hope it is clear that this question cannot be answered in the abstract. How satisfied one is with large classes depends on the consequences one thinks are important. How much does students' sense of engagement matter? Are students' feelings about the course and toward the department important? Is there concern about the percentage of students who go on to enroll in subsequent mathematics courses? The conclusions that one might draw regarding the utility of large classes could vary substantially, depending on how much weight these outcomes are given.

Similar issues arise even if one focuses solely on the mathematics being taught. Suppose one wants to address the question, Do students learn as much mathematics in large classes as in small classes? One must immediately ask, "What counts as mathematics? How much weight will be placed (say) on problem solving, on modeling, or on the ability to communicate mathematically?" Judgments concerning the effectiveness of one form of instruction over another will depend on the answers to these questions. To put things bluntly, a researcher has to know what to look for and what to take as evidence of it before being able to determine whether it is there.

The fact that one's judgments reflect one's values also applies to questions of the type, Which approach works better (or best)? This may seem obvious, but often it is not. Consider calculus reform. Soon after the Tulane "Lean and Lively" conference, whose proceedings appeared in Douglas [5], the National Science Foundation (NSF) funded a major calculus reform initiative. By the mid-1990s NSF program officers were convinced that calculus reform was a "good thing" and that it should be a model for reform in other content areas. NSF brought together mathematicians who had been involved in reform with researchers in mathematics education and posed the following question: "Can we obtain evidence that calculus reform worked (that is, that reform calculus is better than the traditional calculus)?" What they had in mind, basically, was some form of test. They thought it should be easy to construct a test, administer it, and show that reform students did better.

Those who advocated this approach failed to understand that what they proposed would in essence be a comparison of apples and oranges. If one gave a traditional test that leaned heavily on the ability to perform symbolic manipulations, "reform" students would be at a disadvantage because they had not practiced computational skills. If one gave a test that was technology-dependent or that had a heavy modeling component, traditional students would be at a disadvantage because technology and modeling had not been a large part of their curriculum. Either way, giving a test and comparing scores would be unfair. The appropriate way to proceed was to look at the curriculum, identifying important topics and specifying what it means to have a conceptual understanding of them. With this kind of information, individual institutions and departments (and the profession as a whole, if it wished) could then decide which aspects of understanding were most important, which they wanted to assess, and how. As a result of extended discussions, the NSF effort evolved from one that focused on documenting the effects of calculus reform to one that focused on developing a

framework for looking at the effects of calculus instruction. The result of these efforts was the 1997 book *Student Assessment in Calculus* [10].

In sum, many of the questions that would seem natural to ask—questions of the type, What works? or Which method works best?—cannot be answered, for good reason.

Given this, what kinds of questions *can* research in mathematics education address? I would argue that some of the fundamental contributions from research in mathematics education are the following:

- theoretical perspectives for understanding thinking, learning, and teaching;
- descriptions of aspects of cognition (e.g., thinking mathematically; student understandings and misunderstandings of the concepts of function, limit, etc.);
- existence proofs (evidence of cases in which students can learn problem solving, induction, group theory; evidence of the viability of various kinds of instruction);
- descriptions of (positive and negative) consequences of various forms of instruction.

Michèle Artigue's recent *Notices* article [1] describes many of the results of such studies. I will describe some others and comment on the methods for obtaining them in the section "Methods" below.

On Theories and Models (and Criteria for Good Ones)

When mathematicians use the terms "theory" and "models", they typically have very specific kinds of things in mind, both regarding the nature of those entities and the kinds of evidence used to make claims regarding them. The terms "theory" and "models" are sometimes used in different ways in the life sciences and social sciences, and their uses may be more akin to those used in education. In this section I shall briefly walk through the examples indicated in Table 1.

Subject	Mathematics, Physics	Biology	Education, Psychology
Theory of...	Equations; Gravity	Evolution	Mind
Model of...	Heat Flow in a Plate	Predator-Prey Relations	Problem Solving

Table 1. Theories and models in mathematics/physics, biology, and education/psychology.

In mathematics, theories are laid out explicitly, as in the theory of equations or the theory of complex variables. Results are obtained analytically: we prove that the objects in question have the properties we claim they have. In classical physics there is a comparable degree of specificity; physicists specify an inverse-square law for gravitational attraction, for example. Models are understood to be approximations, but they are expected to be very precise approximations in deterministic form. Thus, for example, to model heat flow in a laminar plate, we specify the initial boundary conditions and the conditions of heat flow, and we then solve the relevant equations. In short, there is no ambiguity in the process. Descriptions are explicit, and the standard of correctness is mathematical proof. A theory and models derived from it can be used to make predictions, which in turn are taken as empirical substantiation of the correctness of the theory.

Things are far more complex in the biological sciences. Consider the theory of evolution, for example. Biologists are in general agreement with regard to its essential correctness, but the evidence marshaled in favor of evolution is quite unlike the kind of evidence used in mathematics or physics. There is no way to prove that evolution is correct in a mathematical sense; the arguments that support it consist of (to borrow the title of one of Pólya's books) "patterns of plausible reasoning", along with the careful consideration of alternative hypotheses. In effect, biologists have said the following: "We have mountains of evidence that are consistent with the theory, broadly construed; there is no clear evidence that falsifies the proposed theory, and no rival hypotheses meet the same criteria." While predictions of future events are not feasible given the time scale of evolutionary events, the theory does support an alternative form of prediction. Previously unexamined fossil records must conform to the theory, so that the theory can be used to describe properties that fossils, in particular geological strata, should or should not have. The cumulative record is taken as substantiation for the theory.

In short, theory and supporting evidence can differ substantially in the life sciences and in mathematics and physics. The same holds for models, or at least the degree of precision expected of them: nobody expects animal populations modeled by predator-prey equations to conform to those models in the same way that heat flow in a laminar plate is expected to conform to models of heat flow.

Finally, theories and models in the sciences are always subject to revision and refinement. As glorious and wonderful as Newtonian gravitational theory was, it was superseded by Einstein's theory of relativity. Or consider nuclear theory. Valence theory, based on models of electrons that orbited around nuclei, allowed for amazing predictions, such as the existence of as-yet-undiscovered elements. But physicists no longer talk about electrons in orbit around nuclei; once-solid particles in the theory such as electrons have been replaced in the theory by probabilistic electron clouds. Theories evolve.

Research in mathematics education has many of the attributes of the research in the physical and

life sciences described above. In a "theory of mind", for example, certain assumptions are made about the nature of mental organization—e.g., that there are certain kinds of mental structures that function in particular ways. One such assumption is that there are various kinds of memory, among them working or "short-term" memory. According to the theory, thinking gets done using working memory: that is, the "objects of thought" that people manipulate mentally are temporarily stored in working memory. What makes things interesting (and scientific) is that the theory also places rather strong limits on working memory: it has been claimed (e.g., in [8]) that people can keep no more than about nine "chunks" of information in working memory at one time.

To see that this claim might actually be true, one could try to multiply 379 by 658 with eyes closed. Most people will find it difficult if not impossible. (In a recent meeting I gave a group of about seventy-five mathematicians this task. None of them succeeded within a few minutes.) The reason is that the number of things a person has to keep track of—the original numbers and the various subtotals that arise in doing the multiplication—exceeds nine. Now, a person is better able to do the task mentally after rehearsing some of the subtotals: e.g., a person can calculate $8 \times 379 = 3032$ and repeat "3032" mentally until it becomes a chunk and occupies only one space (a "buffer") in working memory. That leaves enough working space to do other computations. By using this kind of chunking, people can transcend the limits of working memory.[1]

Now consider the truth status of the assertion that people's working memory has no more than about nine slots. There will never be an absolute proof of this assertion. First, it is unlikely that the researchers will find the physical location of working memory buffers in the brain even if they exist; the buffers are components of models, and they are not necessarily physical objects. Second, the evidence in favor of this assertion is compelling but cannot be definitive. Many kinds of experiments have been performed in which people are given tasks that call for using more than nine slots in working memory, and people have failed at them (or, after some effort, performed them by doing what could be regarded as some form of chunking).

As with evolution, there are mountains of evidence that are consistent with this assertion, there is no clear evidence to contradict it, and no rival hypothesis meets the same criteria. But is it proven? No, not in the mathematical sense. The relevant standard is, in essence, what a neutral jury would consider to be evidence beyond a reasonable doubt. The same holds for models of, say, problem solving or (my current interest) models of teaching (see [12], [13]). I am currently engaged in trying to construct a theoretical description that explains how and why teachers do what they do, on the fly, in the classroom. This work, elaborated at the same level of detail as a theory of memory, is called a "theory of teaching-in-context". The claim is that with the theory and with enough time to model a particular teacher, one can build a description of that person's teaching that characterizes his or her classroom behavior with remarkable precision. When one looks at this work, one cannot expect to find the kind of precision found in modeling heat flow in a laminar plate. But (see, e.g., [12]) it is not unreasonable to expect that such behavior can be modeled with the same degree of fidelity to "real-world" behavior as with predator-prey models.

We pursue the question of standards for judging theories, models, and results in the section after next.

Methods

In this article I cannot provide even a beginning catalogue of methods of research in undergraduate mathematics education. As an indication of the magnitude of that task, consider the fact that the *Handbook of Qualitative Research in Education* [6] is nearly 900 pages long! Chapters in that volume include extensive discussions of ethnography (how does one understand the "culture of the classroom", for example?), discourse analysis (what patterns can be seen in the careful study of conversations?), the role of culture in shaping cognition, and issues of subjectivity and validity. And that is qualitative work alone—there is, of course, a long-standing quantitative tradition of research in the social sciences as well. My goal, rather, is to provide an orientation to the kinds of work that are done and to suggest the kinds of findings (and limitations thereof) that they can produce.

Those who are new to educational research tend to think in terms of standard experimental studies, which involve experimental and control groups

[1] *People use "chunking" as a mechanism all the time. A trivial example: one can recall 10-digit phone numbers in part by memorizing 3-digit area codes as a unit. More substantially, the theory asserts that chunking is the primary mechanism that allows one to read this article. Each of the words a person reads is a chunk, which was once a collection of letters that had to be sounded out. The same is the case for all sorts of mathematical concepts that a person now "brings to mind" as a unit. Finally, are "lightning calculators"—the people who do extraordinary mental computations rapidly—a counterexample to the claim made here? It does not appear to be the case. Those who have been studied turn out to have memorized a huge number of intermediary results. For example, many people will bring* "72" *to mind automatically as a chunk when working on a calculation that includes* 9×8*; the "lightning calculators" may do the same for the products of 2- or 3-digit numbers. This reduces the load on working memory.*

and the use of statistics to determine whether or not the results are significant. As it turns out, the use of statistics in education is a much more complex issue than one might think.

For some years from mid-century onward, research in the social sciences (in the United States, at least) was dominated by the example of agriculture. The basic notion was that if two fields of a particular crop were treated identically except for one variable, then differences in crop yield could be attributed to the difference in that variable. Surely, people believed, one could do the same in education. If one wanted to prove that a new way of teaching X was superior, then one could conduct an experiment in which two groups of students studied X—one group taught the standard way, one taught the new way. If students taught the new way did better, one had evidence of the superiority of the instructional method.

Put aside for the moment the issues raised in the previous section about the goals of instruction and the fact that the old and new instruction might not focus on the same things. Imagine that one could construct a test fair to both old and new instruction. And suppose that students were randomly assigned to experimental and control groups, so that standard experimental procedures were followed. Nonetheless, there would still be serious potential problems. If different teachers taught the two groups of students, any differences in outcome might be attributable to differences in teaching. But even with the same teacher, there can be myriad differences. There might be a difference in energy or commitment: teaching the "same old stuff" is not the same as trying out new ideas. Or students in one group might know they are getting something new and experimental. This alone might result in significant differences. (There is a large literature showing that if people feel that changes are made in their own best interests, they will work harder and do better—no matter what the changes actually are. The effects of these changes fade with time.) Or the students might resent being experimented upon.

Here is a case in point. Some years ago I developed a set of stand-alone instructional materials for calculus. Colleagues at another university agreed to have their students use them. In all but two sections the students who were given the materials did better than students who were not given them. However, in two sections there were essentially no differences in performance. It turns out that most of the faculty had given the materials a favorable introduction, suggesting to the students that they would be helpful. The instructor of the sections that showed no differences had handed them out saying, "They asked me to give these to you. I don't know if they're any good."

In short, the classical experimental method can be problematic in educational research. To mention just two difficulties, double blind experiments in the medical sense (in which neither the doctors nor the patients know who is getting the real treatment and who is getting a placebo treatment) are rarely blind, and many experimental variables are rarely controllable in any rigorous sense. (That was the point of the example in the previous paragraph.) As a result, both positive and negative results can be difficult to interpret. This is not to say that such studies are not useful or that large-scale statistical work is not valuable—it clearly is—but that it must be done with great care and that results and claims must be interpreted with equal care. Statistical work of consistent value tends to be that which

a) produces general findings about a population. For example, Artigue [1] notes that "[m]ore than 40% of students entering French universities consider that if two numbers A and B are closer than $1/N$ for every positive N, then they are not necessarily equal, just infinitely close."
b) provides a clear comparison of two or more populations. For example, the results of the Third International Mathematics and Science Study document the baseline performance of students in various nations on a range of mathematical content.
c) provides substantiation, over time, of findings that were first uncovered in more small-scale observational studies.

What one finds for the most part is that research methods in undergraduate mathematics education—in all of education for that matter—are suggestive of results and that the combined evidence of many studies over time is what lends substantiation to findings.

I shall expand on this point with one extended example drawn from my own work. The issue concerns "metacognitive behavior", or metacognition: specifically, the effective use of one's resources (including time) during problem solving.

Here is a motivating example. Many years ago, when one standard first-year calculus topic was techniques of integration, the following exercise was the first problem on a test given to a large lecture class:

$$\int \frac{x}{x^2 - 9}\, dx.$$

The expectation was that the students would make the obvious substitution $u = (x^2 - 9)$ and solve the problem in short order. About half the class did. However, about a quarter of the class, noticing that the denominator was factorable, tried to solve the problem using the technique of partial fractions. Moreover, about 10 percent of the students, noticing that the denominator was of the form $(x^2 - a^2)$, tried to solve the problem using the substitution $x = 3 \sin \theta$. All of these methods yield the correct answer, of course, but the second and third are very time consuming for students. The students who

used those techniques did poorly on the test, largely because they ran out of time.

Examples such as this led me to develop some instructional materials that focused on the strategic choices that one makes while working integration problems. The materials made a difference in student performance. This provided some evidence that strategic choices during problem solving are important.

The issue of strategic choices appeared once again when, as part of my research on problem solving, I examined videotapes of students trying to solve problems. Quite often, it seemed, students would read a problem statement, choose a solution method quickly, and then doggedly pursue that approach even when the approach did not seem to be yielding results. To make such observations rigorous, I developed a "coding scheme" for analyzing videotapes of problem solving. This analytical framework provided a mechanism for identifying times during a problem session when decision making could shape the success or failure of the attempt. The framework was defined in such a way that other researchers could use it, not only for purposes of examining my tapes, but also for examining their own as well. Using it, researchers could see how students' decision making helped or hindered their attempts at problem solving.

Such frameworks serve multiple purposes. First, having such a scheme allows the characterization of videotapes to become relatively objective: if two trained analysts working on the same tape independently produce the same coding of it, then there is reason to believe in the consistency of the interpretation. Second, having an analytic tool of this type allows one to trace the effects of problem-solving instruction: "before and after" comparisons of videotapes of problem-solving sessions can reveal whether students have become more efficient or effective problem solvers. Third, this kind of tool allows for accumulating data across studies. The one-line summary of results in this case: metacognitive competence is a very productive factor in problem solving.[2] For extensive detail, see [9].

As indicated above, research results in education are not "proven" in the sense that they are proven in mathematics. Moreover, it is often difficult to employ straightforward experimental or statistical methods of the type used in the physical sciences because of complexities related to what it means for educational conditions to be "replicable". In education one finds a wide range of research methods. A look at one of the first volumes on undergraduate mathematics education, namely [14], suggests the range. If anything, the number and type of methods have increased, as evidenced in the three volumes of *Research in Collegiate Mathematics Education*. One finds, for example, reports of detailed interviews with students, comparisons of reform and traditional calculus, an examination of calculus "workshops", and an extended study of one student's developing understanding of a physical device and graphs related to it. Studies employing anthropological observation techniques and other qualitative methods are increasingly common.

How valid are such studies, and how much can we depend on the results in them? That issue is pursued immediately below.

[2] *In the case at hand (metacognitive behavior), a large number of studies have indicated that effective decision making during problem solving does not "come naturally". Such skills can be learned, although intensive instruction is necessary. When students learn such skills, their problem-solving performance improves.*

Standards for Judging Theories, Models, and Results

There is a wide range of results and methods in mathematics education. A major question then is the following: How much faith should one have in any particular result? What constitutes solid reason, what constitutes "proof beyond a reasonable doubt"?

The following list puts forth a set of criteria that can be used for evaluating models and theories (and more generally any empirical or theoretical work) in mathematics education:

- Descriptive power
- Explanatory power
- Scope
- Predictive power
- Rigor and specificity
- Falsifiability
- Replicability
- Multiple sources of evidence ("triangulation")

I shall briefly describe each.

Descriptive Power

By descriptive power I mean the capacity of a theory to capture "what counts" in ways that seem faithful to the phenomena being described. As Gaea Leinhardt [7] has pointed out, the phrase "consider a spherical cow" might be appropriate when physicists are considering the cow in terms of its gravitational mass—but not if one is exploring some of the cow's physiological properties! Theories of mind, problem solving, or teaching should include relevant and important aspects of thinking, problem solving, and teaching respectively. At a very broad level, fair questions to ask are: Is anything missing? Do the elements of the theory correspond to things that seem reasonable? For example, say a problem-solving session, an interview, or a classroom lesson was videotaped. Would a person who read the analysis and saw the videotape reasonably be surprised by things that were missing from the analysis?

Explanatory Power

By explanatory power I mean providing explanations of how and why things work. It is one thing to say that people will or will not be able to do

certain kinds of tasks or even to describe what they do on a blow-by-blow basis; it is quite another thing to explain why. It is one thing, for example, to say that people will have difficulty multiplying two three-digit numbers in their heads. But that does not provide information about how and why the difficulties occur. The full theoretical description of working memory, which was mentioned above, comes with a description of memory buffers, a detailed explanation of the mechanism of chunking, and the careful delineation of how the components of memory interact with each other. The explanation works at a level of mechanism: it says in reasonably precise terms what the objects in the theory are, how they are related, and why some things will be possible and some not.

Scope

By scope I mean the range of phenomena covered by the theory. A theory of equations is not very impressive if it deals only with linear equations. Likewise, a theory of teaching is not very impressive if it covers only straight lectures!

Predictive Power

The role of prediction is obvious: one test of any theory is whether it can specify some results in advance of their taking place. Again, it is good to keep things like the theory of evolution in mind as a model. Predictions in education and psychology are not often of the type made in physics.

Sometimes it is possible to make precise predictions. For example, Brown and Burton [4] studied the kinds of incorrect understandings that students develop when learning the standard U.S. algorithm for base 10 subtraction. They hypothesized very specific mental constructions on the part of students—the idea being that students did not simply fail to master the standard algorithm, but rather that students often developed one of a large class of incorrect variants of the algorithm and applied it consistently. Brown and Burton developed a simple diagnostic test with the property that a student's pattern of incorrect answers suggested the false algorithm he or she might be using. About half of the time they were then able to predict the incorrect answer that the student would obtain to a new problem before the student worked the problem!

Such fine-grained and consistent predictions on the basis of something as simple as a diagnostic test are extremely rare of course. For example, no theory of teaching can predict precisely what a teacher will do in various circumstances; human behavior is just not that predictable. However, a theory of teaching can work in ways analogous to the theory of evolution. It can suggest constraints and even suggest likely events.

[Making predictions is a very powerful tool in theory refinement. When something is claimed to be impossible and it happens, or when a theory makes repeated claims that something is very likely and it does not occur, then the theory has problems! Thus, engaging in such predictions is an important methodological tool, even when it is understood that precise prediction is impossible.]

Rigor and Specificity

Constructing a theory or a model involves the specification of a set of objects and relationships among them. This set of abstract objects and relationships supposedly corresponds to some set of objects and relationships in the "real world". The relevant questions are:

How well defined are the terms? Would you know one if you saw one? In real life? In the model? How well defined are the relationships among them? And how well do the objects and relations in the model correspond to the things they are supposed to represent? As noted above, one cannot necessarily expect the same kinds of correspondences between parts of the model and real-world objects as in the case of simple physical models. Mental and social constructs such as memory buffers and the "didactical contract" (the idea that teachers and students enter a classroom with implicit understandings regarding the norms for their interactions and that these understandings shape the ways they act) are not inspectable or measurable in the ways that heat flow in a laminar plate is. But we can ask for detail, both in what the objects are and in how they fit together. Are the relationships and changes among them carefully defined, or does "magic happen" somewhere along the way? Here is a rough analogy. For much of the eighteenth century the phlogiston theory of combustion—which posited that in all flammable materials there is a colorless, odorless, weightless, tasteless substance called "phlogiston" liberated during combustion—was widely accepted. (Lavoisier's work on combustion ultimately refuted the theory.) With a little hand waving, the phlogiston theory explained a reasonable range of phenomena. One might have continued using it, just as theorists might have continued building epicycles upon epicycles in a theory of circular orbits.[3] The theory might have continued to produce some useful results, good enough "for all practical purposes". That may be fine for practice, but it is problematic with regard to theory. Just as in the physical sciences, researchers in education have an intellectual obligation to push for greater clarity and specificity and to look for limiting cases or counterexamples to see where the theoretical ideas break down.

Here are two quick examples. First, in my research group's model of the teaching process, we represent aspects of the teacher's knowledge, goals, beliefs, and decision making. Skeptics (including us) should ask, how clear is the representation? Once terms are defined in the model (i.e., once we specify a teacher's knowledge, goals, and beliefs) is there

[3] *This example points to another important criterion,* simplicity. *When a theory requires multiple "fixes" such as epicycles upon epicycles, that is a symptom that something is not right.*

hand waving when we say what the teacher might do in specific circumstances, or is the model well enough defined so that others could run it and make the same predictions? Second, the "APOS theory" expounded in [2] uses terms such as Action, Process, Object, and Schema. Would you know one if you met one? Are they well defined in the model? Are the ways in which they interact or become transformed well specified? In both cases the bottom line issues are, What are the odds that this too is a phlogiston-like theory? Are the people employing the theory constantly testing it in order to find out? Similar questions should be asked about all of the terms used in educational research, e.g., the "didactical contract", "metacognition", "concept image", and "epistemological obstacles".

Falsifiability

The need for falsifiability—for making nontautological claims or predictions whose accuracy can be tested empirically—should be clear at this point. It is a concomitant of the discussion in the previous two subsections. A field makes progress (and guards against tautologies) by putting its ideas on the line.

Replicability

The issue of replicability is also intimately tied to that of rigor and specificity. There are two related sets of issues: (1) Will the same thing happen if the circumstances are repeated? (2) Will others, once appropriately trained, see the same things in the data? In both cases answering these questions depends on having well-defined procedures and constructs.

The phrasing of (1) is deliberately vague, because it is supposed to cover a wide range of cases. In the case of short-term memory, the claim is that people will run into difficulty if memory tasks require the use of more than nine short-term memory buffers. In the case of sociological analyses of the classroom, the claim is that once the didactical contract is understood, the actions of the students and teacher will be seen to conform to that (usually tacit) understanding. In the case of beliefs, the claim is that students who hold certain beliefs will act in certain ways while doing mathematics. In the case of epistemological obstacles or APOS theory, the claims are similarly made that students who have (or have not) made particular mental constructions will (or will not) be able to do certain things.

In all of these cases the usefulness of the findings, the accuracy of the claims, and the ability to falsify or replicate depend on the specificity with which terms are defined. Consider this case in point from the classical education literature. Ausubel's theory of "advance organizers" in [3] postulates that if students are given an introduction to materials they are to read that orients them to what is to follow, their reading comprehension will improve significantly. After a decade or two and many, many studies, the literature on the topic was inconclusive: about half of the studies showed that advance organizers made a difference, about half not. A closer look revealed the reason: the very term was ill defined. Various experimenters made up their own advance organizers based on what they thought they should be—and there was huge variation. No wonder the findings were inconclusive! (One standard technique for dealing with issues of well-definedness, and which addresses issue (2) above, is to have independent researchers go through the same body of data and then compare their results. There are standard norms in the field for "inter-rater reliability"; these norms quantify the degree to which independent analysts are seeing the same things in the data.)

Multiple Sources of Evidence ("Triangulation")

Here we find one of the major differences between mathematics and the social sciences. In mathematics one compelling line of argument (a proof) is enough: validity is established. In education and the social sciences we are generally in the business of looking for *compelling evidence*. The fact is, evidence can be misleading: what we think is general may in fact be an artifact or a function of circumstances rather than a general phenomenon.

Here is one example. Some years ago I made a series of videotapes of college students working on the problem, How many cells are there in an average-size human adult body? Their behavior was striking. A number of students made wild guesses about the order of magnitude of the dimensions of a cell—from "let's say a cell is an angstrom unit on a side" to "say a cell is a cube that's 1/100 of an inch wide." Then, having dispatched with cell size in seconds, they spent a very long time on body size, often breaking the body into a collection of cylinders, cones, and spheres and computing the volume of each with some care. This was *very* odd.

Some time later I started videotaping students working problems in pairs rather than by themselves. I never again saw the kind of behavior described above. It turns out that when they were working alone, the students felt under tremendous pressure. They knew that a mathematics professor would be looking over their work. Under the circumstances they felt they needed to do *something* mathematical, and volume computations at least made it look as if they were doing mathematics! When students worked in pairs, they started off by saying something like "This sure is a weird problem." That was enough to dissipate some of the pressure, the result being that there was no need for them to engage in volume computations to relieve it. In short, some very consistent behavior was actually a function of circumstances rather than being inherent in the problem or the students.

One way to check for artifactual behavior is to vary the circumstances: to ask, do you see the same thing at different times in different places? Another

is to seek as many sources of information as possible about the phenomenon in question and to see whether they portray a consistent message. In my research group's work on modeling teaching, for example, we draw inferences about the teacher's behavior from videotapes of the teacher in action—but we also conduct interviews with the teacher, review his or her lesson plans and class notes, and discuss our tentative findings with the teacher. In this way we look for convergence of the data. The more independent sources of confirmation there are, the more robust a finding is likely to be.

Conclusion

The main point of this article has been that research in (undergraduate) mathematics education is a *very* different enterprise from research in mathematics and that an understanding of the differences is essential if one is to appreciate (or better yet, contribute to) work in the field. Findings are rarely definitive; they are usually suggestive. Evidence is not on the order of proof, but is cumulative, moving towards conclusions that can be considered to be beyond a reasonable doubt. A scientific approach is possible, but one must take care not to be *scientistic*—what counts are not the trappings of science, such as the experimental method, but the use of careful reasoning and standards of evidence, employing a wide variety of methods appropriate for the tasks at hand.

It is worth remembering how young mathematics education is as a field. Mathematicians are used to measuring mathematical lineage in centuries, if not millennia; in contrast, the lineage of research in mathematics education (especially undergraduate mathematics education) is measured in decades. The journal *Educational Studies in Mathematics* dates to the 1960s. The first issue of Volume 1 of the *Journal for Research in Mathematics Education* was published in January 1970. The series of volumes *Research in Collegiate Mathematics Education*—the first set of volumes devoted solely to mathematics education at the college level—began to appear in 1994. It is no accident that the vast majority of articles cited by Artigue [1] in her 1999 review of research findings were written in the 1990s; there was little at the undergraduate level before then! There has been an extraordinary amount of progress in recent years, but the field is still very young, and there is a very long way to go.

Because of the nature of the field, it is appropriate to adjust one's stance toward the work and its utility. Mathematicians approaching this work should be open to a wide variety of ideas, understanding that the methods and perspectives to which they are accustomed do not apply to educational research in straightforward ways. They should not look for definitive answers but for ideas they can use. At the same time, all consumers and practitioners of research in (undergraduate) mathematics education should be healthy skeptics. In particular, because there are no definitive answers, one should certainly be wary of anyone who offers them. More generally, the main goal for the decades to come is to continue building a corpus of theory and methods that will allow research in mathematics education to become an ever more robust basic and applied field.

References

[1] M. Artigue, The teaching and learning of mathematics at the university level: Crucial questions for contemporary research in education, *Notices Amer. Math. Soc.* **46** (1999), 1377–1385.

[2] M. Asiala, A. Brown, D. de Vries, E. Dubinsky, D. Mathews, and K. Thomas, A framework for research and curriculum development in undergraduate mathematics education, *Research in Collegiate Mathematics Education* (J. Kaput, A. Schoenfeld, and E. Dubinsky, eds.), vol. II, Conference Board of the Mathematical Sciences, Washington, DC, pp. 1–32.

[3] D. P. Ausubel, *Educational Psychology: A Cognitive View*, Holt-Reinhardt-Winston, New York, 1968.

[4] J. S. Brown and R. R. Burton, Diagnostic models for procedural bugs in basic mathematical skills, *Cognitive Science* **2** (1978), 155–192.

[5] R. G. Douglas (ed.), *Toward a Lean and Lively Calculus*, MAA Notes Number 6, Mathematical Association of America, Washington, DC, 1986.

[6] M. LeCompte, W. Millroy, and J. Preissle, *Handbook of Qualitative Research in Education*, Academic Press, New York, 1992.

[7] G. Leinhardt, On the messiness of overlapping goals in real settings, *Issues in Education* **4** (1998), 125–132.

[8] G. Miller, The magic number seven, plus or minus two: Some limits on our capacity for processing information, *Psychological Review* **63** (1956), 81–97.

[9] A. H. Schoenfeld, *Mathematical Problem Solving*, Academic Press, Orlando, FL, 1985.

[10] ____ (ed.), *Student Assessment in Calculus*, MAA Notes Number 43, Mathematical Association of America, Washington, DC, 1997.

[11] ____ , On theory and models: The case of teaching-in-context, *Proceedings of the XX Annual Meeting of the International Group for Psychology and Mathematics Education* (Sarah B. Berenson, ed.), Psychology and Mathematics Education, Raleigh, NC, 1998.

[12] ____ , Toward a theory of teaching-in-context, *Issues in Education* **4** (1998), 1–94.

[13] ____ , Models of the teaching process, *Journal of Mathematical Behavior* (in press).

[14] D. Tall (ed.), *Advanced Mathematical Thinking*, Kluwer, Dordrecht, 1991.