19 Research methods in (mathematics) education

Alan H. Schoenfeld

University of California – Berkeley

Though this be madness, yet there is method in't.

Hamlet, Act II, Scene 2, Line 205

I. INTRODUCTION

Four hundred years post-*Hamlet*, researchers in mathematics education might well invert Polonius's famous comment: in contemporary research, *though this be method, yet there is madness in't* seems fairly close to the mark. In this chapter I will unravel some of the reasons for the madness and the method in contemporary research, suggest criteria regarding the appropriateness and adequacy of investigatory methods and their theoretical underpinnings, and identify some productive pathways for the development of beginning researchers' skills and understandings.

Let us begin with an indication of the magnitude of the task. In terms of scale, it is worth noting that the *Handbook of Qualitative Research in Education* (LeCompte, Millroy, & Preissle, 1992) and the *Handbook of Research Design in Mathematics and Science Education* (Kelley & Lesh, 2000) are 881 and 993 pages long, respectively. Neither of these volumes claims to cover the territory; moreover, there is relatively little overlap between them. The vast majority of the former is devoted to ethnographic research, which is but one of many approaches to understanding what happens in educational settings. The core of the latter is devoted to elaborating half a dozen "research designs that are intended to radically increase the relevance of research to practice" such as teaching experiments and computer-modeling studies. Neither discusses the kinds of quantitative or experimental methods that dominated educational research just a few decades ago (see, e.g., Campbell & Stanley, 1966), and which remain important to understand—increasingly so in the United States, where the federal government is placing great stress on the "gold standard" randomized controlled trials (see, e.g., Whitehurst, 2002, 2003). Given the broad spectrum of contemporary work, any attempt in a single chapter to deal with educational research methods must of necessity be selective rather than comprehensive. Hence, I have adopted the strategy of identifying and elucidating major themes. For an accessible introduction to a broad range of methods, readers might wish to examine Green, Camilli, and Elmore (2006).

Second, some historical perspective is in order. Mathematics education is solidly grounded in psychology and philosophy among other fields, and can thus claim to have a long and distinguished lineage. However, the discipline of research in mathematics education is itself quite young. The first meeting of the International Congress on Mathematics Education was held in 1968. Volume 1 of *Educational Studies in Mathematics* appeared in May 1968. The *Zentralblatt für Didaktik der Mathematik* was first published in June 1969, the *Journal for Research in Mathematics Education* in January 1970. In addition, while growth in the field has been substantial, that growth has been anything but evolutionary. As a consequence first of the "cognitive revolution" in the 1970s and 1980s and then of an expanded emphasis on sociocultural issues and methods in the 1990s the field has, within its short life span, completely reconceptualized the nature of the phenomena considered to be central, and it has developed new methods to explore them. Although significant progress has been made, educational research has hardly entered a period of "normal science." What we take to be foundational assumptions, how we investigate various empirical phenomena, and how we provide warrants for the claims we make, are all issues that stand in need of clarification and elaboration. To the degree that space allows, those issues will be addressed in this chapter.

Sections II and III of this chapter provide the broad context for the discussions that follow. Section II provides a brief summary of trends in mathematics education over the 20th century, describing the philosophical underpinnings and research methods of some major approaches to the study of mathematical thinking, teaching, and learning. A main point is that mathematics education research is a young discipline, having coalesced in the last third of the century. This serves, in large measure, to explain the diversity of perspectives and methods; some degree of chaos is hardly surprising during the early stages of a discipline's formation. Section III summarizes the current state of affairs, with an eye toward the future. An argument is made that the "pure *versus* applied" characterization of much research may be a misdirection—that educational research has progressed to the point where it can address many basic issues in the context of meaningful applications.

The core of this chapter, Sections IV, V, and VI, is devoted to the elaboration of a framework that addresses the purposes and conduct of research. It addresses the role of underlying assumptions in the conduct of research, the implications of (implicit or explicit) choices of theoretical frameworks and methods for the quality of the research findings, and the nature of the warrants one can make regarding research findings.

As a rough heuristic guide for the discussion of methods, the following framework is used. Research contributions will be conceptualized along three dimensions: their *trustworthiness* (how much faith can one put in any particular research claim?), their *generality* (are claims being made about a specific context, a well defined range of contexts, or are they supposedly universal?), and their *importance.* Section IV provides an overarching description of the research process: the choice of conceptual framework, the focal choice of data and their representation, their analysis, and the interpretation of the analyses. This description focuses largely on places where essential decisions are made, and on possible problems regarding trustworthiness and generality when such decisions are made. Section V offers a set of criteria by which educational theories, models, and findings can be judged. Section VI elaborates on the framework discussed above, with an emphasis on the *generality* dimension of the framework. It offers a series of examples illustrating the kinds of claims that can be made, ranging from those that make no claim of generality ("this is what happened here"), to those that claim to be universal ("the mind works in the following ways"). For each class of examples discussed, issues of trustworthiness and importance are addressed.

Section VII addresses issues related to the preparation of educational researchers. The educational enterprise is complex and deeply interconnected, and simple approaches to simple problems are not likely to provide much purchase on the major issues faced by the field. But, beginning researchers have to start somewhere. Is there a reasonable way to bootstrap into the necessary complexity? Are there general chains of inquiry, and pathways into educational research, that seem promising or productive? Brief concluding remarks are made in Section VIII.

II. A BRIEF HISTORY: PERSPECTIVES AND METHODS

Throughout much of the 20th century, a range of perspectives and their associated research methods competed for primacy in mathematics education. Some of those perspectives were: associationism/behaviorism, Gestaltism, constructivism, and later, cognitive science and sociocultural theory.

Associationism and behaviorism were grounded in the common assumption that learning is largely a matter of habit formation, the consistent association of particular stimuli in an organism's environment with particular events or responses. The generic example of behaviorism is that of Pavlov's dogs, which salivated at the sound of their handlers' approach—the association between certain noises and their upcoming meals being so strong that it induced a physiological response. Pavlov showed experimentally that the response could be reinforced so strongly that the dogs salivated in response to stimulus noises even when no food was present. Presumably, human learning was similar. E. L. Thorndike's 1922 volume *The Psychology of Arithmetic* established the foundation for pedagogical research and practices grounded in associationism. Thorndike's learning theory was based on the concept of mental *bonds*, associations between sets of stimuli and the responses to them (for example, "three times five" and "fifteen"). Like muscles, bonds became stronger if exercised and tended to decay if not exercised. Thorndike proposed that, in instruction, bonds that "go together" should be taught together. This theoretical rationale provided the basis for extended repetitions (otherwise known as "drill and practice") as the vehicle for learning.

In broad-brush terms, associationist/behaviorist perspectives held sway at the beginning of the 20th century—at least in the United States. Evidence thereof may be found in two yearbooks, the very first *Yearbook* of the (U.S.) National Council of Teachers of Mathematics, was published in 1926, and the 1930 *Yearbook* of the (U.S.) National Society for the Study of Education (NSSE).

A theory based on the development of bonds and associations lends itself nicely to empirical research. From the associationist perspective, a fundamental goal is to develop sequences of instruction that allow students to master mathematical procedures efficiently, with a minimum of errors. Thus, relevant research questions pertain to the nature of drill – how much, and of what type. Such work was relatively new, heralding the beginnings of a "scientific" approach to mathematics instruction. It is interesting to note, for example, that the editors of NCTM's 1926 *Yearbook, A General Survey of Progress in the Last Twenty-Five Years,* introduced a research chapter (Clapp, 1926) with the following statement:

> Detailed investigations and controlled experiments are distinctly the product of the last quarter century. The Yearbook would not be truly representative of the newest developments without a sampling of the newer types of materials that are developing to guide our practice. (NCTM, 1926/1995, p. 166)

The 1926 NCTM Yearbook contains two chapters that focus on research. The first (Schorling, 1926) invokes Thorndike and provides an extensive summary of "The psychology of drill in mathematics" (pp. 94–99), including a list of twenty "principles which have been of practical help to [the author] in the organization and administration of drill materials." The second, mentioned above (Clapp, 1926), represented the state of the art in the study of student learning of arithmetic. An empirical question, for example, was to determine which arithmetic sums students find more difficult. Clapp reports:

> In the study of the number combinations a total of 10,945 pupils were tested. The number of answers to combinations was 3,862,332…. [The sums] were read to pupils at the approximate rate of one combination every two seconds. The rate was determined by experimentation and the time was made short enough to prevent a pupil's counting or getting the answer in any other round-about way…. The purpose of the study was to determine which combinations had been reduced to the automatic level. The results may be said to indicate the relative learning difficulty of the combinations. (Clapp, 1926, pp. 167–168)

The 1930 NSSE *Yearbook* (Whipple, 1930), was devoted to the study of mathematics education. Its underpinnings were avowedly behaviorist/associationist:

> Theoretically, the main psychological basis is a behavioristic one, viewing skills and habits as fabrics of connections. This is in contrast, on the one hand, to the older structural psychology which still has to make direct contributions to classroom procedure, and on the other hand, to the more recent *Gestalt* psychology, which, though promising, is not yet ready to function as a basis of elementary education. (Knight, 1930, p. 5)

Thus, in the 1930 NSSE *Yearbook* one saw research studies examining the role of drill in the learning of multiplication (Norem & Knight, 1930) and fractions (Brueckner & Kelley, 1930), and on the effectiveness of mixed drill in comparison to isolated drill (Repp, 1930). Errors were studied in fine-grained detail, similarly to the work reported by Clapp. In their study of multiplication, for example, Norem and Knight (1930) analyzed the patterns found in 5365 errors made by students practicing their multiplication tables.

It is interesting to note from the perspective of these *Yearbook* authors and editors, Gestalt psychology, while "promising," was not ready for prime time with regard to mathematics instruction. In many ways, the Gestaltists' stance could be seen as antithetical to that of the associationists:

> With the development of "field theories" of learning, of which the Gestalt theory is most familiar to school teachers, the center of interest shifted from what was often, and perhaps unjustly, called an "atomistic" concept of learning to one which emphasized understanding of the number system and number relations and which stressed problem solving more than drill on number facts and processes. (Buswell, 1951, p. 146)

Indeed, insight and structure were central concerns of the Gestaltists. An archetypal Gestaltist story is Poincaré's (1913; see also Hadamard, 1945) description of his discovery of the structure of Fuchsian functions. Poincaré describes having struggled with the problem for some time, then deliberately putting it out of mind and taking a day trip. He reports that as he boarded a bus for an excursion, he had an inspiration regarding the solution, which he verified upon his return.

Poincaré's story is typical, both in substance and methods. With regard to substance, the outline of the story is the basic tale of Gestalt discovery: one works as hard as possible on a problem, lets it incubate in the subconscious, has an insight, and verifies it. Similar stories are told concerning the chemist Kekulé's dreaming of a snake biting its tail, and realizing that benzene must be ring-like in structure, and of Archimedes (in the bath) solving the problem of how to determine whether King Heron's crown is pure gold, without damaging the crown itself. With regard to method, what Poincaré offers is a retrospective report.

Perhaps the best known advocate of the Gestaltist perspective with regard to schooling is Max Wertheimer. Wertheimer's (1945) classic book, *Productive Thinking,* is a manifesto against "blind drill" and its consequences. Wertheimer describes the responses he obtained from students when he asked them to work problems such as

$$\frac{357+357+357}{3}=?$$

He reported that some "bright subjects" saw through such problems, observing that the division "undoes" the addition, yielding the original number. Wertheimer found, however, that these students were the exception rather than the rule. Many students who had earned high marks in school were blind to the structure of the problem, and insisted on working through it mechanically. He continues,

> These experiences reminded me of a number of more serious experiences in schools, which had worried me. I now looked more thoroughly into customary methods, the ways of teaching arithmetic, the textbooks, the specific psychology books on which their methods were based. One reason for the difficulty became clearer and clearer: the emphasis on mechanical drill, on 'instantaneous response,' on developing blind, piecemeal habits. Repetition is useful, but continuous use of mechanical repetition also has harmful effects. It is dangerous because it easily induces habits of sheer mechanized action, blindness, tendencies to perform slavishly instead of thinking, instead of facing a problem freely. (Wertheimer, 1945, pp. 130–131)

The focus of the Gestaltists' work—whether in discussions of schooling or in discussions of professionals' mathematical and scientific thinking (e.g., in Poincaré's story and Wertheimer's interviews of Einstein regarding the development of the theory of relativity), was on meaning, on insight, on structure. Their methods were "introspectionist," depending on individuals' reports of their own thinking processes. These methods, alas, proved unreliable. As Peters (1965) summarizes subsequent research, "a wealth of experimental material [demonstrated] the detailed effects of attitudes and interests on what is perceived and remembered. Perception and remembering are now regarded as processes of selecting, grouping, and reconstructing. The old picture of the mind as receiving, combining, and reproducing has finally been abandoned" (p. 694). And, one might add, methods that depended on individuals' reports of their own mental processes could hardly be depended upon.

Following World War II, the "scientific" approach to research in education returned with a vengeance. Given the context, this was natural. It was science that had brought an end to the war, and it was science that promised a brighter future. (The motto of one major corporation, for example, was "progress is our most important product." There is no doubt that the progress referred to was scientific.) After a decade of worldwide turmoil, what could be more psychologically desirable—or prestigious—than the prospect of a rational, orderly way of conducting one's business?

The wish to adopt the trappings of science played out in various ways. Among them were the ascendancy of the radical behaviorists and the dominance of "experimental" methods in education—and more broadly in the social sciences, so named for the reasons discussed in the previous paragraph. First, consider the radical behaviorists. As noted above, the "mentalistic" approaches of groups such as the Gestaltists, depending on introspection and retrospective reports, were shown to be unreliable. The radical behaviorists such as B.F. Skinner (see, e.g., Skinner, 1958) took this objection to reports of mental processes one step further. They declared that the very notions of "mind" and "mental structures" were artifactual and theoretically superfluous; all that counted was (observable and thus quantifiable) behavior.

The radical behaviorists, following in the tradition of their classical behaviorist antecedents, took much of their inspiration and methods from studies of animals. Rats and pigeons might not be able to provide retrospective verbal reports, but they could learn—and their behaviors could be observed and tallied. One could teach a pigeon to move through a very complicated sequence of steps, one step at a time, by providing rewards for the first step until it became habitual, then the second step after the first, and so on. Out of such work came applications to human learning. Resnick (1983) describes Skinner's approach as follows:

> [Skinner] and his associates showed that "errorless learning" was possible through shaping of behavior by small successive approximation. This led naturally to an interest in a technology of teaching by organizing practice into carefully arranged sequences through which the individual gradually acquires the elements of new and complex performance without making wrong responses en route. This was translated for school use into "programmed instruction"—a form of instruction characterized by very small steps, heavy

> prompting, and careful sequencing so that children could be led step by step toward ability to perform the specified behavioral objectives. (Resnick, 1983, pp. 7–8)

It is worth noting that the holy grail of "errorless learning" persisted long after the behaviorists' day in the sunshine had supposedly passed. Many well-known computer-based tutoring systems marched students through various procedures one step at a time, refusing to accept as correct inputs that, even if ultimately sensible, were "errors" in the sense that they were not the "most logical" input anticipated by the program.

Behaviorism, both in its earlier and then in its radical form, was one manifestation of scientism in the research culture. As indicated above, scientism was widespread, permeating all of the social sciences during the third quarter of the 20th century. It played out in the wholesale and ofttimes inappropriate adoption of statistical and "experimental" methods through much of the third quarter of the century. Many educational experiments were modeled on the "treatment A *versus* treatment B" model used in agricultural or medical research.

Under the right conditions, comparison studies can provide tremendously useful information. If, for example, two fields of some crop are treated almost identically and there is a significant difference in yield between them, that difference could presumably be attributed to the difference in treatment (which might be the amount of watering, the choice of fertilizer, etc.). Drug tests operate similarly, with "experimental" and "control" groups being given different treatments. Statistical analyses indicate whether the treatment drug has significantly different effects than the control (typically a placebo).

Unfortunately, the "right conditions" rarely held in the educational work described above. Although it may be possible to control for all but a few variables in agricultural research, the same is not the case for most educational comparisons. If different teachers taught "experimental" and "control" classes, the "teacher variable" might be the most significant factor in the experience. Or, the same teacher might teach the two treatments at different times of the day, and the fact that one group met early in the morning and the other right after lunch (or the teacher's enthusiasm for one treatment) would make a difference. Or, the "treatment" itself might be ill defined. There are many ways in which ostensibly straightforward experimental comparisons can go wrong.

Half a dozen years ago, when I wrote the draft of this chapter for the first edition of this *Handbook,* I wrote the following: "One could say much more, but there is no need to flog a dead horse; by and large, the field has abandoned such methods. This in itself is unfortunate; the use of quantitative methods may need to be revisited (with care)" (Schoenfeld, 2002b, p. 240). It is ironic that in the United States there has been a resurgence of experimentalism—but in many ways without the care and attention that needs to be devoted to experimental methods, and thus with the serious possibility of a return to the scientism of the mid-1900s. We return to the issue of experimental methods at the end of this section and in Section IV.

Let us continue the historical narrative. Slowly, and in various ways, U.S. mathematics education researchers made their way out of the paradigmatic and methodological straightjackets of the 1960s and 1970s. In many ways, work had simply run itself into the ground, and the field came to recognize that fact. For example Kilpatrick discussed the methodological state of the art in the mid-1970s as follows:

> No one is suggesting that researchers abandon the designs and techniques that have served so well in empirical research. But a broader conception of research is needed …
>
> Some years ago a group of researchers gave a battery of psychological tests each summer to mathematically talented senior high school students.… The scores on the tests were intercorrelated, and some correlation coefficients were significant, some not. Several research reports were published.… As Krutetskii (1976) notes, the process of solution did not appear to interest the researchers—yet what rich material could have been obtained from these gifted students if one were to study their thinking processes in dealing with

> mathematical problems. Why were the students simply given a battery of tests to take instead of being asked to solve mathematical problems? It's a good question. (Kilpatrick, 1978, p. 18)

In a hugely ironic twist, the study of mind was largely resuscitated by artificial intelligence (AI), the study of "machines that think." Pioneering efforts in AI included computer programs such as Newell and Simon's (1972) "General Problem Solver," or GPS. GPS played a reasonable game of chess. It solved "cryptarithmetic" problems.[1] And it solved problems in symbolic logic. Specifically, GPS derived 51 of the first 53 results in Russell and Whitehead's famous mathematical treatise *Principia Mathematica*—and GPS's proof of one result was shorter than the proof provided by Russell and Whitehead.

In order to write problem solving programs, Newell and Simon asked people to solve large numbers of problems, working on them "out loud" so that researchers could record and later analyze what was done as their subjects worked on the problems. They transcribed the recordings and pored over the transcripts, looking for productive patterns of behavior—that is, for strategies that mimicked the successful "moves" made by their subjects. Those strategies, once observed and abstracted, were then written up as computer programs.

The irony comes from the fact that AI provided the means for hoisting the behaviorists by their own dogmatic petard. AI programs "worked"—their record of problem solving was clear. More importantly, all of their workings were out in the open—every decision was overtly specified. By virtue of being inspectable and specifiable, work in AI met all of the behaviorists' criteria for being scientific. At the same time, the success of the AI enterprise depended entirely on the investigation of human thought processes. Hence, the study of "mind" was legitimized. Studies of human thinking and research methods that involved reports of problem solving were once again scientifically "acceptable."

Given the climate at the time, the legitimization of such methods was by no means easy. There was a great deal of controversy over the use of problem solving protocols (records of out-loud problem solutions), and it took some years before the dust settled (see, e.g., Ericsson & Simon, 1980; Nisbett & Wilson, 1977). At the same time, a wide variety of methods and perspectives became known internationally, sowing the seeds for the profusion of views and techniques that would flower in the latter part of the century. Piaget had, of course, been developing a massive corpus of work on children's intellectual development, both philosophical (e.g., Piaget, 1970) and with regard to various mathematical concepts such as number, time, and space (e.g., Piaget, 1956, 1969a, 1969b). Piaget's work brought the "clinical interview" to prominence as a research method. The work of Krutetskii (1976) and colleagues (see Kilpatrick & Wirszup, 1975) popularized the idea of "teaching experiments," detailed studies of principled attempts at instruction and their consequences in terms of students' abilities to engage with mathematics. Freudenthal's (1973, 1983) work lay the foundation for the study of "realistic mathematics," a central tenet of which is that mathematics instructional sequences should be grounded in contexts and experiences that support the development of meaningful mathematical abstractions.

Broadly speaking, the 1970s and the 1980s were a time of explosive growth. The "cognitive revolution" (see, e.g., Gardner, 1987) brought with it a significant epistemological shift, and with it, new classes of phenomena for investigation and new methods for exploring them. For much of the century, the focus of research in mathematical thinking and learning had been on *knowledge*—a body of facts and procedures to be mastered. As theoretical frameworks evolved, such knowledge was seen to be only one (albeit very important) aspect of mathematical thinking. Theoretical frameworks (see, e.g., Schoenfeld, 1985) indicated that central aspects of mathematical performance included the knowledge base, problem solving strategies, aspects of metacognition, and beliefs. They invoked the notion of "culture," in that students were seen to engage in (often counterproductive) practices derived from their experiences in school, and which were quite different from the practices of the mathematical community. Each of these aspects of cognition was explored with a wide variety of emerging methods: observational and experimental studies, teaching experiments, clinical interviews, the analysis of "out loud" protocols, computer modeling, and more.

There were, however, few ground rules for conducting such research—either in terms of investigatory norms or in terms of quality standards. Research in mathematics education had moved from a period of "normal science" to one in which the ground rules were unknown. Fundamental questions, not well addressed, became: How does one define new phenomena of interest? How does one look for them, document them? How does one make sense of things such as the impact of metacognitive decision making on problem solving performance, the relationship between culture and cognition, or what might be an appropriate focus for investigation in the blooming complexity of a teaching experiment? The field began to address such issues: see, for example, Schoenfeld, 1992.

The flowering of theoretical perspectives and methods continued through the end of the 20th century. Specifically, sociocultural perspectives had long roots. Vygotsky, for example,[2] in both *Mind in Society* (1978) and also *Thought and Language* (1962) had advanced a perspective that, perhaps in too-simple terms, could be seen as complementary to Piaget's. Vygotsky and his theoretical allies argued that learning is a function of social interaction:

> Human learning presupposes a specific social nature and a process by which children grow into the intellectual life of those around them. (Vygotsky, 1978, p. 88)

> Every function in the child's social development occurs twice: first, on the social level, and later on the individual level; first, *between* people (*interpsychological)*, and then inside the child (*intrapsychological*). This applies equally to voluntary attention, to logical memory, and to the formation of concepts. All the higher functions originate as actual relations between human individuals. (Vygotsky, 1978, p. 57)

From the 1970s onward, multiple lines of research explored aspects of cognition and culture. For example, a series of studies conducted in Brazil (see Carraher, 1991 for a review) explored the relationships between mathematical understandings in school and in "real world" contexts such as candy selling. The main theoretical perspective adopted by the French for studies of mathematical didactics presumed the existence of a "didactical contract" that is inherently social in nature (see, e.g., Brousseau, 1997). German work, significantly shaped by Bauersfeld (e.g., 1980, 1993), took as a given that there are multiple realities and social agendas playing out in instruction, and that one must attend to "language games" (à la Wittgenstein) in the mathematics classroom. By the time of ICME VII in Quebec (1992), a multiplicity of competing theoretical perspectives had blossomed. The Proceedings of the VII International Congress on Mathematics Education's Working Group on Theories of Learning (Steffe, Nesher, Cobb, Goldin, & Greer, 1996), for example, contains three large sections: "Sociological and anthropological perspectives on mathematics learning," "Cognitive science theories and their contributions to the learning of mathematics," and "The contributions of constructivism to the learning of mathematics," as well as a fourth small section that includes explorations of metaphor as the possible basis for a theory of learning of mathematics (see also English, 1997; Sfard, 1994).

At the end of the previous century and the beginning of the present one, one saw a proliferation of perspectives, of theories, and of methods. On the one hand, this is undoubtedly healthy: the field had escaped from the paradigmatic and theoretical straightjackets of the earlier part of the 20th century, and it was virtually bursting with energy and excitement. To give but two seminal examples, the year 2002 saw the publication of two journal special issues that served both to problematize the state of research and to move it forward. A special issue of *Mathematical Thinking and Learning* (Nasir & Cobb, 2002) offered theoretical reframings of issues related to diversity, equity, and mathematical learning. And, a special issue of the *Journal of the Learning Sciences* (Sfard & McClain, 2002) presented a series of papers, all of which analyzed a common body of video data. Questions of what one can say, with what assurance, about a body of data, will help the field to move forward. Reflective analysis is entirely appropriate for a young field that is very clearly not in a time of "normal science" (Kuhn, 1970).

Since the publication of the first edition of this *Handbook,* there has been a significant shift in the climate surrounding educational research in the Unites States. The U.S. Department of Education, specifically the Institute for Education Sciences (IES), has taken a strong stance in favor of applied evaluative research employing the "gold standard" of randomized controlled trials (see, e.g., Whitehurst, 2002, 2003). IES has made a major investment in the creation of the What Works Clearinghouse (WWC; see http://www.whatworks.ed.gov/), whose task it is to conduct reviews of the literature in search of studies that meet rigorous analytic standards; it has also funded the creation of a new educational research society, the Society for Research on Educational Effectiveness (see http://www.sree-net.org/), whose purpose it is to "to advance and disseminate research on the causal effects of education interventions, practices, programs, and policies." What will come of these efforts remains to be seen. As noted above, the use of experimental methods should indeed be revisited: quantitative methods have been under-utilized in the past few decades, and such methods should be part of the researcher's tool kit. However, beginning efforts are hardly auspicious. I was WWC's first Senior Content Advisor for its mathematics studies. I resigned in 2005 when WWC failed to address some fundamental flaws in its published study reports, and then took actions that resulted in the cancellation of the special journal issue in which my discussion of these issues was to appear. A case can be made that there is now in education an unusually strong and unfortunate mixing of political and intellectual agendas, much as there is in the case of the suppression of research on global warming that is contrary to current federal ideology. Details may be found in Schoenfeld (2006, March).

III. THE CURRENT STATE OF AFFAIRS[3]

This section focuses on current needs. It begins with some assertions about desiderata for research, and moves on to a discussion of current challenges. It takes as background the current and somewhat chaotic state of research: that there are multiple and competing theoretical perspectives, and a host of methods that are tailored to specific problems, but of limited general utility.

The first two assertions regarding desiderata for high-quality research live in dialectic tension:

1. One must guard against the dangers of compartmentalization. It is all too easy to focus narrowly, ignoring or dismissing work or perspectives not obviously related to one's own. This can be costly, given the systemic and deeply connected nature of educational phenomena. Educators need a sense of the "big picture" and of how things fit together.
2. One must guard against the dangers of being superficial. Superficial knowledge (of information or methods) is likely to yield trivial research. Generally speaking, high quality research comes when one has a deep and focused understanding of the area being examined, and extended experience mulling over the issues under question.

Needless to say, it is difficult to strive toward both depth and breadth simultaneously. Yet they are both necessary.

The third assertion has to do with the current state of theory and methods:

3. Educational research has reached the point where it is possible to conduct meaningful research in contexts that "matter," and not simply in the laboratory. Indeed, the traditional model of doing basic research and then applying it in context needs to be reconsidered.

The fourth and fifth assertions deal with the conduct of research, and serve as announcements of themes to be elaborated at length in this chapter:

4. In conducting research, one must have a sense of where one stands and where one thinks he or she is heading. On stance: One has biases and a theoretical perspective, whether one thinks so or not. These affect what one "sees." On direction: Methods are not used in the abstract, or pulled off the shelf. Any research method is, in effect, a lens or filter through which phenomena are viewed (and possibly clarified, or distorted, or obscured). Thus, the ways that questions are framed should shape the ways that methods are selected and employed.
5. A fundamental issue both for individuals conducting their own studies and for the field as a whole is the need to develop a deep understanding of what it means to make and justify claims about educational phenomena. What is a defensible claim? What is the scope of that claim? What kinds of evidence can be taken as a legitimate warrant for that claim?

The discussion that follows provides some brief examples in support of the first two claims. The third assertion is elaborated at greater length, for it characterizes the evolving contexts within which it is now possible to do basic research and to "make a difference." The fourth and fifth assertions point to the main substance of this chapter. A few points will be made briefly, by way of orientation to what follows.

Issue 1: The dangers of compartmentalization

It is easy to find examples of the costs of educational myopia. One can point fingers at those administrators who fail to value subject matter understanding, such as those who want to know how (not *if*) they can "retrain" surplus social studies teachers, in the break between school years, to teach mathematics the following year. Or, one can point to an inverse problem, mathematicians who believe that knowing the subject matter is all that is required by way of the preparation of mathematics teachers. Indeed, major educational movements have been ill conceived and squandered large amounts of money, precisely because of their tunnel vision. Consider, for example, various attempts to provide students with preparation, in school, for productive work lives. This is known as the "school-to-work" problem.

Most proposed school-to-work programs are oriented toward *skills*: They may proceed by examining the workplace, identifying productive skills, and teaching them directly. Or, they may take a more "contextual" approach, with suggestions that students should engage in apprenticeships and/or that curricula should be designed to reflect workplace demands. Such approaches are doomed to fail. There is, of course, a pragmatic reason. The skills set is a moving target, in that skills learned today will be obsolete tomorrow, and new skills will be needed. More importantly, there is a deep theoretical reason. The past quarter century of research in mathematics education has shown that skills are but one component of mathematical performance. Problem solving strategies, metacognition, beliefs, and domain-specific practices are also aspects of mathematical behavior. These are essential components of a theory of mathematical behavior. And it's not just mathematics; a good case can be made that they are relevant in any domain.

If you want people to be good at X, then you ought to have a theory of what it means to be good at X. A starting place for the dimensions of such a theory should be established theories of competence from other areas. For example, school-to-work attempts that proceed in ignorance of theoretical frameworks and advances in mathematics education—ignoring, for example, crucial aspects of performance such as metacognition and beliefs—proceed at their own peril. Similarly, mathematics educators must ask, what frameworks or insights from other fields (for example, anthropology, or studies of organizational behavior) does mathematics education ignore at its own peril? There is, thus, the need for great breadth.

Issue 2: The dangers of superficiality

Breadth may be essential, as argued immediately above, but there is a significant breadth-versus-depth tradeoff. Expertise comes with focus (which takes time and energy), and the danger of a lack of focus is either dilettantism or superficiality. A case in point is discussed by Heath (1999), a linguist who points to the problem of researchers in education using methods from other fields without really understanding them. For example, Heath discusses "discourse analysis"—which, in educational research, often seems to mean "making sense of what people say by whatever ad hoc methods seem appropriate." Heath notes that there are more than a half dozen different schools of discourse analysis, each with its own traditions, history, and methods. Moreover, each type makes particular kinds of contributions, takes a fair amount of time to master, and should not be used cavalierly by amateurs. Educational researchers who call their ad hoc attempts to make sense of dialogue "discourse analysis" are abusing the term. That, of course, is but one example. In mathematics education, one could point to the cavalier use of "clinical interviews," "protocol analysis," and other methods.

Do we need such methods? The answer is a clear yes—more so as time goes on. But those who employ such methods need to be well enough steeped in them to use them with wisdom, and skill. This implies focus and depth.

Issue 3: The relationship between research and practice

In *Pasteur's Quadrant: Basic Science and Technical Innovation*, Donald Stokes (1997) discusses tensions between theory and applications in science and technology. Stokes argues that in both our folk and scientific cultures, basic and applied research are viewed as being in tension. For most, applications would seem the *raison d'être* of science. Stokes points out, however, that in elite circles, "pure" science has been considered far superior to its applications. A quote from C. P. Snow's essay on "the two cultures" describes how scientists at Cambridge felt about their work: "We prided ourselves that the science that we were doing could not, in any conceivable circumstances, have any practical use. The more firmly one could make the claim, the more superior one felt." (Recall G. H. Hardy's famous (1967) quote: "I have never done anything 'useful'.... The case for my life [is] … that I have added something to knowledge.")

This perspective was reified by Vannevar Bush, who was asked by President Franklin Delano Roosevelt to map out a plan for post-World War II scientific research and development. Bush's report, *Science, the Endless Frontier,* ultimately provided the philosophical underpinnings of the U. S. National Science Foundation (NSF).

Echoing Snow, Bush wrote that "basic research is performed without thought of practical ends" and that its defining characteristic is "its contribution to 'general' knowledge and an understanding of nature and its laws." He went on to say that if one tries to mix basic and applied work, that "applied research invariably drives out pure." Federal funding should support basic work, he argued; out of that basic work would come a broad range of applications. The tension between basic and applied work is represented in Figure 19.1.

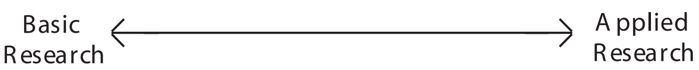


Figure 19.1 Basic and applied research seen as polar opposites.

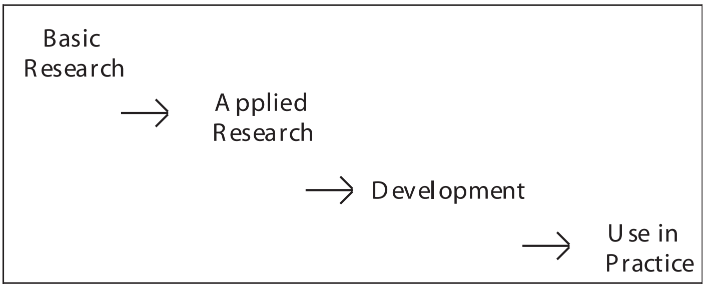


Figure 19.2 The progression from research to use (after Stokes, 1997, p. 10).

A hypothesized progression from basic research to use-in-practice is represented in Figure 19.2.

One can point to researchers whose work fits cleanly at various points in the spectrum illustrated in Figure 19.1—paradigmatic examples being Niels Bohr and Thomas Edison. Bohr's work on the structure of the atom was conducted without thought of applications; in contrast, Edison disdained theory while pouring his energy into "electrifying" the United States. You can imagine Bohr situated on the far left of Figure 19.1, and Edison on the far right.

But what about Louis Pasteur?

Pasteur's work in elaborating biological mechanisms at the microbiological level—working out the "germ theory of disease"—is as basic as you can get. But, Pasteur did not engage in this activity solely for reasons of abstract intellectual interest. He was motivated by problems of spoilage in beer, wine, and milk, and the hope of preventing and/or curing diseases such as anthrax, cholera, rabies, and tuberculosis.

At what point on the spectrum in Figure 19.1 should one place Pasteur? Do we split him in half, with 50% at each end of the spectrum? Or should one "average" his contributions, placing him in the middle? Neither does him justice.

Stokes resolves this dilemma by disentangling these two aspects of Pasteur's work, considering basic knowledge and utility as separate dimensions of research. He offers the scheme given in Figure 19.3.

Pasteur has a home in this scheme—and moreover, *considerations of use* and *the quest for fundamental understanding* are seen as living in potential synergy. Note that this conceptualization destroys the linearity of the hypothetical scheme in Figure 19.2: there are times when

Research is inspired by:

		Considerations of Use?	
		No	Yes
Quest for Fundamental Understanding?	Yes	Pure Basic Research (Bohr)	Use-Inspired Basic Research (Pasteur)
	No		Pure Applied Research (Edison)

Figure 19.3 A two-dimensional representation of "basic" and "applied" considerations for research (Stokes, 1977, p. 73).

"basic" work can be done (indeed, may need to be done) in applied contexts. Fundamental research does not necessarily take place before, or in contexts apart from, those of practical use.

This perspective, elaborated by Stokes in the case of science, applies equally well to educational research. The idea, simply stated, is that a significant proportion of educational research can now be carried out in "real" contexts.[4] The careful study of "design experiments" or other educational interventions can reveal important basic information, about mathematical thinking, teaching, and learning.

This statement, however, raises profound questions for the conduct of research. Issues of how to make sense of "real world phenomena," and how to justify the claims one makes, are thorny indeed. Mathematics education in particular, and educational research in general, have yet to grapple adequately with methods and standards for making and judging such claims.

Issue 4: On knowing where you are and where you are going

One point that is made repeatedly in this chapter is that, whether or not researchers believe that they have theoretical perspectives and biases, they do. (Researchers who think otherwise are like the proverbial fish who are unaware of the medium in which they swim.) This observation is critically important, for one's framing assumptions shape what one will attend to in research. They also, needless to say, affect the scope and robustness of one's findings. Sections IV and V of this chapter will address these issues at some length.

The second point that needs to be made here is more subtle, and is easily misinterpreted. That is: research methods are best chosen when one has some idea of what it is one is looking for. A research method is a lens through which some set of phenomena is viewed. A lens may bring some phenomena sharply into focus. But it may also blur others at the same time, and perhaps even create artifactual or illusory images. Moreover, to continue the metaphor, different lenses are appropriate for different purposes—the same individual may use one set of glasses for close-up work, one for regular distance, and complex devices such as telescopes for very long-distance work. So it is with methods: the phenomena we wish to "see" should affect our choice of method, and the choice of method will, in turn, affect what we are capable of seeing. And, of course, the kinds of claims one will be able to make (convincingly) will depend very much on the methods that have been employed.

Thus, researchers should be very much aware of the following questions, and the answers they propose to them:

- What theoretical perspective undergirds the work?
- What questions are being asked? What kinds of claims does one expect to make?
- What methods are appropriate to address these questions?
- What kinds of warrants do these methods provide in substantiation of the (potential) claims to be made?

These questions are essentially independent of the nature of one's work—that is, they apply equally well to naturalistic research intended to provide "rich, thick descriptions," to experimental methods employing statistical analyses, or to the construction of "models" representing various phenomena. If the researcher does not have good answers to them, there is a good chance the research will be seriously flawed.

It is essential to stress that not all decisions about methods must be made beforehand; the claims above are not intended to be either reductivist or positivist. Research is a dialectic process in which researchers come to grips with phenomena by living with them, and understandings evolve over time. One can point to numerous studies in which important phenomena emerged mid-stream. Indeed, longstanding notions such as "grounded theory" and methodological tools such as the "constant comparative method" (see Glaser & Strauss, 1967) serve as codifications of the fact that sense-making is an inductive process. The same is true of work that includes significant quantitative components: in "design experiments" (see, e.g., Brown, 1992) and various teaching interventions (see, e.g., Ball & Lampert, 1999; Schauble & Glaser, 1996), a great deal of data are gathered, and then sifting and winnowing processes takes place. What is essential to understand, however, is that the sifting and winnowing are done with the purpose of answering specific (perhaps emergent) questions. If the question isn't clear by the end of the process, the answer isn't likely to be either.

Issue 5: What is believable and why?

This, of course, is *the* key question the field faces with regard to methods. It is, alas, all too infrequently addressed. One could hardly hope to answer the question in a chapter of this nature—but, one can hope to bring it to the forefront and clarify aspects of it. Most of the balance of this chapter (Sections IV, V, and VI) is devoted, directly or indirectly, to that enterprise.

IV. A VIEW OF THE RESEARCH PROCESS AND ITS IMPLICATIONS

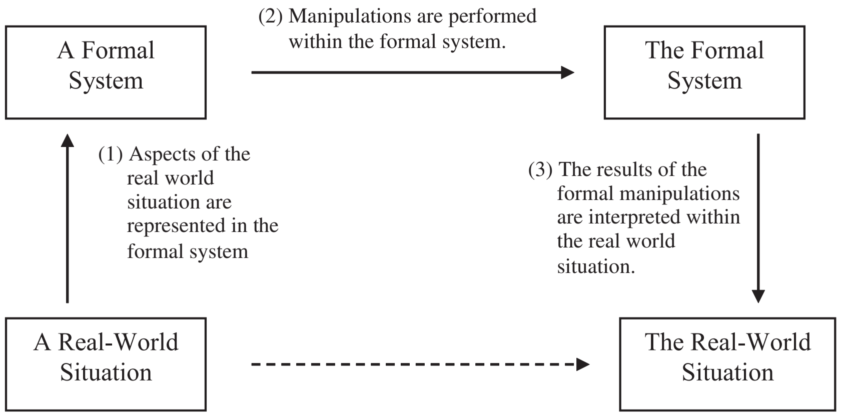
In Section II of this chapter I noted that, especially in the decades following World War II, there was extensive use of experimental methods in education—and afterwards, the recognition that such methods had not produced much of lasting value. Partly as a result of those problems, such methods (modeled on those in the physical and biological sciences) had fallen out of favor; however, the "gold standard" of randomized controlled trials is now being urged upon educational researchers once again. It is worth reconsidering the issue of experimental methods, to better understand why they have contributed so little thus far and to insure that when they are used, they are used correctly. The reasons for doing so are not merely historical (although post mortems often reveal interesting and useful information) or prescriptive. The fact is that many of the problems that plagued experimental studies also have the potential to weaken or negate the value of studies that employ non-experimental research methods. And, those who do not learn from the mistakes of the past are doomed to repeat them.

This section of the chapter begins with a description of a conceptual framework within which one can examine the use of experimental methods. The framework is employed to highlight potential difficulties with such methods—places where the work can be undermined if researchers are not appropriately careful. The framework is then expanded and modified so that it applies to non-experimental methods as well. This will set the stage for later discussions (Section VI) regarding the trustworthiness and robustness of educational research findings.

The use of statistical/experimental methods is a form of modeling. A simple diagram (see Figure 19.4 below) and discussion, taken from Schoenfeld, 1994, highlight some of the issues involved in the use of such models.

It should be noted that statistical tests are conducted under the assumption that the "real world" situations being considered conform to the conditions of specific statistical models; if they do not, the conclusions drawn are invalid. When the experimental conditions do match those of the statistical model, it is then assumed that the results of statistical analyses conducted provide valid interpretations of the real world situations. This is represented by the dashed arrow at the bottom of Figure 19.4.

The essential point to keep in mind when applying statistical models is that they, like any other models, are *representations* of particular situations—and the usefulness of the model will depend on the fidelity of the representation. The effective use of statistical or other modeling techniques to shed light on a real-world situation depends on the accuracy of all three mappings illustrated in Figure 19.4: (1) the abstraction of aspects of the situation into the


A Formal System
(2) Manipulations are performed within the formal system.
The Formal System
(1) Aspects of the real world situation are represented in the formal system
(3) The results of the formal manipulations are interpreted within the real world situation.
A Real-World Situation
The Real-World Situation

Steps 1, 2, and 3 combined yield an analysis of the real-world situation. Note that the quality of the analysis depends on the mappings to and from the formal system (steps 1 and 3). If either one of those mappings is flawed, the analysis is not valid.

Figure 19.4

model, (2) the mathematical analysis within that model, and (3) the mapping of interpretations back into the situation. Even if the manipulations performed within the formal system (e.g., calculations of statistical significance) are correct, there is no guarantee that the interpretation of the results obtained in the formal system will accurately reflect aspects of the real-world system from which the model was abstracted.

There are numerous places where these mappings can break down. For example, statistical significance means nothing if the conditions under which experimentation is done do not conform to the assumptions of the model underlying the development of the statistics; and it means little if the constructs being examined are ill-defined. Although researchers adopted the language of "treatments" and "variables," the objects they so named often failed to have the requisite properties: ofttimes, for example, an instructional "treatment" was not a univalent entity but was very different in the hands of two different experimenters or teachers. Similarly, if an instructional experiment used different teachers for the treatment and control groups, then teacher variation (rather than the instructional treatments) might account for observed differences; if the same teacher taught both groups, there still might be a difference in enthusiasm, or in student selection. In short, many factors other than the ones in the statistical model—the variables of record—could and often did account for important aspects of the situation being modeled (Schoenfeld, 1994, pp. 700–701).

With some expansion, the scheme given in Figure 19.4 can be modified into a scheme that applies to *all* observational and experimental work, whether that work is qualitative or quantitative in nature. See Figure 19.5.[5]

The first main change from Figure 19.4 to Figure 19.5 is the explicit recognition (seen along the vertical dimension) that "reality" is never abstracted directly. There is, of course, the fact that humans do not perceive reality directly: we interpret our sensory images of the world through conscious or unconscious filtering mechanisms. More to the point, however, is the fact that any act of codifying (our perceptions of) the real world represents an act of selection, and thus of theoretical commitment. Whatever perspective the analyst adopts, some things are highlighted and some are downplayed or ignored. This set of choices—the set of entities and relationships selected for analysis—will be called the analyst's *conceptual model.* It indicates what "counts," from the analyst's perspective.

The second main change is the expansion from experimental methods to general analytic methods. When one employs classical statistical/experimental methods, one typically performs

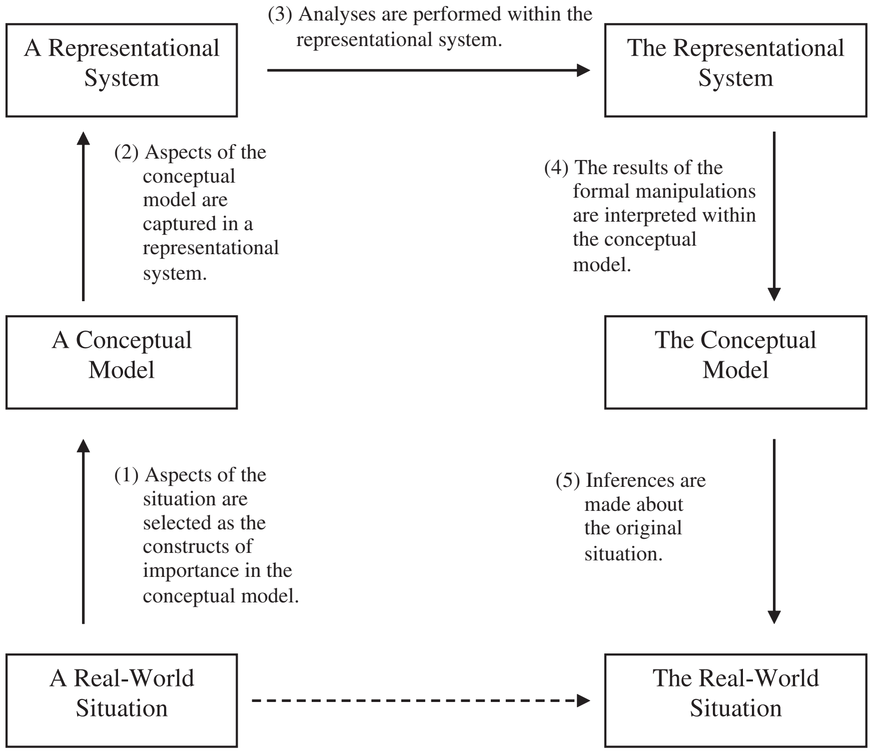

A Representational System
(3) Analyses are performed within the representational system.
The Representational System
(2) Aspects of the conceptual model are captured in a representational system.
(4) The results of the formal manipulations are interpreted within the conceptual model.
A Conceptual Model
The Conceptual Model
(1) Aspects of the situation are selected as the constructs of importance in the conceptual model.
(5) Inferences are made about the original situation.
A Real-World Situation
The Real-World Situation

Figure 19.5 A schematic representation of the analytic process in general.

standard statistical manipulations (t- or z-tests, factor analyses, etc.) under the assumption that the data gathered conform to the conditions of some formal statistical model. In "treatment A *versus* treatment B" comparison tests, for example, one gathers relevant data such as the scores of the two treatment groups on some outcome measure. These are data points in a formal statistical model (the upper left-hand corner of Figure 19.4). The data are analyzed in accord with the conditions of the model (arrow 2 in Figure 19.4). If the results are deemed statistically significant, the researcher typically draws the inference that the (significant) difference in performance can be attributed to the difference in the two treatments.

That situation can be abstracted as follows (see the top section of Figure 19.5). Virtually any record of occurrences can be considered "data." (The reliability and utility of such data are an issue, of course.) Such records may, for example, be field notes of anthropological observations; audio- or videotapes of a classroom, or of one or more people engaged in problem solving; interview transcripts, or just about any permanent record of events. This record (which already represents a filtering of events through the researcher's conceptual lens) is then represented in some way for purposes of analysis. Videotapes may be coded for gestures, or for the nature and kind of interactions between people. Field notes and interview transcripts may be annotated and categorized. Then, the analyst pores over the data. There are myriad ways to do so, of course. Coded data may be analyzed statistically, as in the "process-product" paradigm, where the independent variables were typically tallies of specific classroom behaviors by teachers and dependent variables were measures of student performance on various tests (see, e.g., Brophy & Good, 1986; Evertson, Emmer, & Brophy, 1980). Statistical analyses may be integrated with observational data, as in *The AAUW Report: How Schools Shortchange Girls* (1992) or Boaler (2002). Patterns of behavior may be "captured" in a model, as in Schoenfeld

(1998). Or, patterns of observations and other data may be woven together in a narrative, as in Eisenhart, Borko, Underhill, Brown, Jones, and Agard (1993). No matter what the form, the point is that inferences are drawn, within a conceptual framework, on the basis of data captured in a representational system. Following such analyses, the researchers map their findings back to (their interpretation of) the phenomena they are investigating.

Seen from this perspective, all instances of interpretation and analysis—quantitative and qualitative alike—are seen to be similar in some fundamental ways. The interpretive pathway begins with the (conscious or unconscious) imposition of an interpretive framework. It continues with the selection and representation of data considered relevant to the question at hand, and the interpretation of those data within the conceptual and representational framework. The interpretation is then mapped back to the real world, as an explanation/interpretation of the phenomena at hand.

Having established this general framework, I would now like to point to some of the difficulties involved in traversing the pathways indicated in Figure 19.5. Those difficulties point to potential pitfalls in quantitative and qualitative studies alike.

Along the first arrow: Focal choices reflect (perhaps implicit) theoretical commitments

The first and perhaps most fundamental point that must be recognized in the conduct of educational research is that what researchers see in complex real world settings is not objective reality but a complex function of their beliefs and understandings. In any setting, infinitely many things might catch one's interest. Where one's attention settles is shaped by what one believes is important and by what one is prepared to see. Take, for example, the question of "what counts" in mathematical understanding. For the behaviorist/associationist, the central issue in mathematics learning is the efficacy with which one masters standard procedures; understanding was defined to mean "performing the procedures well." From this point of view, Clapp's 1926 study of students performing a total of 3,862,332 arithmetic sums, described in the previous section, makes perfect sense. Indeed, from the radical behaviorist's point of view, any invocation of mental processes above and beyond the strength of "bonds" developed by repeated practice was nonsense. In contrast, the Gestaltist Wertheimer looked for signs of structural understanding. Where traditionalists saw arithmetic competence, Wertheimer (1945) saw "blind, piecemeal habits … tendencies to perform slavishly instead of thinking." Or, consider an event such as an hour's mathematics lesson. From the "process-product" perspective (see, e.g., Brophy & Good, 1986), what counts are teacher behaviors and student performance on various outcome measures; other aspects of classroom interactions might be ignored. From the situative perspective, a central issue may be the joint construction of mathematical meaning in the classroom, through discourse (Greeno et al., 1998). From a cultural perspective, one might focus on the typicality of certain instructional practices within and across nations (Stigler & Hiebert, 1999). From a microsociological perspective (see, e.g., Bauersfeld, 1995), one might focus on the nature of the *classroom culture* and the role of language in the classroom as a "medium between person and world." Researchers viewing the same phenomena from within these different traditions might attend to, and "see," very different things. The analytic frames they then construct—the conceptual models in Figure 19.5—will then differ widely. See, for example, Sfard & McClain, 2002: different researchers see very different focal phenomena in the same video record of a classroom. See also Schoenfeld (in press), in which there are four complementary analyses of a classroom interaction. One analysis focuses on the nature of the work a teacher must do to orchestrate productive mathematics learning; one on teachers' decision making; one on a particularly productive discourse structure, and one on equity issues. A major analytic challenge is to build frameworks that allow researchers to converge on similar or synergistic findings when focusing on such data (see, e.g., Sherin & Sherin, in press).

Along the second arrow: Do the data, as represented, reflect the constructs of importance in the conceptual model?

Perhaps the most accessible cases in point regarding this issue come from the statistical and experimental paradigms. As noted in Section II, "treatment A versus treatment B" comparisons are meaningless if the treatments are ill-defined (e.g., the case of "advance organizers"), or if the ostensible "independent variables" in an experiment are not the only variables that affect performance on the outcome variables. Moreover, very different interpretations can result depending on what one takes as relevant. How, for example, does one represent "mathematical proficiency" or "subject matter knowledge"? If one study comparing two curricula uses a test of basic skills as its outcome measure and another study uses a test that focuses on conceptual understanding and problem solving as its outcome measure, the studies could well report (apparently) contradictory findings.

The question of data representation is central in all paradigms. To start with an obvious point, field notes are clearly selective—they represent the observers' focus and biases. Less obvious but equally important, "objective" records such as videotapes also represent observers' focus and biases. If there is one camera, where is it focused? On an individual, on a group? On the written work produced by individuals or groups, or on their faces as they talk? (With more than one camera, one can get more "coverage," but issues of focus remain.) Given a videotape record, which occurrences get coded—and at what grain size—when the tape is analyzed? Compare, for example, the fine level of detail in the transcript coding scheme set forth by Lucas, Branca, Goldberg, Kantowsky, Kellogg, and Smith (1980) with the rather coarse-grained coding scheme found in Schoenfeld (1985). Both schemes were aimed at understanding "problem solving." Yet, by their very nature, they supported very different kinds of analyses. Or, compare the two transcripts of "Leona's puppy story" by Sarah Michaels (pp. 241–244) and James Gee (pp. 244–246) in their discussion of discourse analysis (Gee, Michaels, & O'Connor, 1992). Michaels presents the story in narrative form, with a range of markers to indicate changes in pitch and intonation, timing, and more. False starts and repairs are included, in an attempt to capture a large part of the "spoken record" in written notation. In contrast, Gee strips such markers from the text. He presents a cleaned-up version in "stanzas," as a narrative poem—an "ideal realization" of the text. Here too (and this is the point of the authors' examples), two different transcripts of the same oral record support two very different types of analyses. The form of representation makes a difference.

In short, the constructs in the conceptual model may or may not be well defined, and the ways data are represented may or may not correspond in straightforward ways to those constructs. To pick an example from the social realm: "turn-taking" is easy to code, but coding "mathematical authority" and "social authority" (see, e.g., Cobb, 1995) is a much more delicate issue. The choices of what to code, and the accuracy, consistency, and grain size of the coding, will have a critical impact on the quality of the analysis.

Along the third arrow: What's meaningful within the representational scheme? What ban be said about the quality of the inferences drawn?

The third arrow asks the deceptively simple question, "What conclusions can be drawn within the given conceptual system, using the data as represented?"

Putting technical language aside for the moment, there are a series of common-sense questions that are natural to ask when someone proposes to make some judgments from a body of data. Those include the following:

- Are there enough data on which to base a solid judgment?
- Is the means of analysis consistent—that is., will anyone trained in the analytical methods draw the same conclusions from the same data?

- Does the data-gathering mechanism tap into stable phenomena—that is., will someone be likely to produce similar data when assessed at different times, and will their interpretation be consistent?

In terms of classical statistical methods, these questions are related to technical issues of sampling, reliability, and validity. There is, of course, a huge body of statistical and psychometric theory and technique that addresses those issues. Unfortunately, however, the theoretical underpinnings and the conditions of application for those theories and techniques mesh very poorly with evolving epistemological understandings regarding *theories of competence* in subject matter domains. In days gone by, tests such as the U.S. National Assessment of Educational Progress (NAEP) simply used "content by difficulty matrices," where test items reflected mastery of particular topics at various levels of difficulty. Currently, the situation is very much more complex. Theories of mathematical understanding include aspects of competence such as the ability to employ problem solving strategies, to employ self-regulatory skills effectively, and more. "Performance assessment" items may cross topic areas—a problem may be accessible to a solution via algebraic or geometric means, for example, or be solvable numerically or symbolically. Under such circumstances, standard psychometric techniques are woefully inadequate to provide knowledge profiles of students. New methods need to be developed (see, e.g., Glaser & Linn, 1997; Greeno, Pearson, & Schoenfeld, 1997).

In terms of the broad spectrum of research methods available to (mathematics) educators today, the questions highlighted above are both fundamental and extremely difficult. Section V of this chapter is devoted to addressing such issues.

Along the fourth arrow: Are results derived in the representational system meaningful in the conceptual model?

Arrow 4 is the mirror image of arrow 2, completing the analytic loop within the conceptual model—the pathway from arrow 2 through arrow 4 represents the gathering, analysis, and interpretation of data, given the assumptions of the conceptual model. The main point here is that, no matter how fine the analysis within the representational system may be, the overall analysis is no better than the mapping to and from the conceptual model.

As one case in point, consider econometric analyses of school district expenditures vis-à-vis the effects of class size. Ofttimes precise data regarding actual class size are unavailable. In early studies, researchers used proxies for these input data—for example, the ratio of "instructional staff" to students in a district, or some fraction thereof. But such ratios had the potential to be tremendously misleading, because some districts' figures included non-teaching administrators and some did not. As a result, the input variables had no consistent meaning. The output variables were often standardized tests of basic skills or other "achievement tests." Typically, these tests were only marginally related to the actual curricula being taught—and thus were dubious measures of the effectiveness of instruction. In short, both input and output variables in many such studies were of questionable value. No matter how perfectly executed the statistical analyses on such data might be, the results are close to meaningless.

Within the standard statistical paradigms, there are also well known examples of "sampling error"—one example of which was a mid-20th century telephone poll of voters in a U.S. presidential election. What the pollsters failed to realize was that telephones were not universal, and that by restricting their sample to people who owned telephones, they had seriously biased the sample (and, indeed, made the wrong prediction). Much more recently, there is the fact that a large number of medical studies were conducted using only male patients. The samples were randomly drawn and the statistics were properly done; the findings applied perfectly well to the male half of the population. The problem is that the findings were also assumed—in many cases incorrectly—to apply to the female half of the population as well.

Issues may be more subtle with regard to qualitative data, but they are there all the same. In the polling example just given, the poll was accurate for the population sampled, but not for the population at large. Similarly, some Piagetian findings, which were once thought to be universal, were later seen to be typical of middle-class Swiss children but not of children who had very different backgrounds. Sampling error is every bit as dangerous a flaw in qualitative as in quantitative research.

The same is the case for issues of construct validity. In Piagetian clinical interviews, for example, children's performance on certain (wonderfully clever) interview tasks was taken as evidence of the presence or absence of certain cognitive structures. Further studies revealed that although performance on certain tasks might be robust, the robustness was in part a function of the task design; other tasks aimed at the same mental constructs did not necessarily produce the same results. In terms of Figure 19.5, the analysis within the representational system (performance on a set of tasks) was just fine, but the mapping back to the conceptual framework (the attribution of certain logico-deductive structures on the basis of the analyses) was questionable. The issues are hardly more straightforward when the constructs involved are things such as "power relationships" and "self-concept."

Another example where construct validity is problematic is that of IQ. If IQ is defined by performance on various IQ tests, one obtains (relatively) consistent scores. But when one thinks of such scores as reflecting "intelligence," one opens a can of worms. Historically speaking, Binet thought of his tests as identifying places where people needed remediation—a somewhat questionable but defensible position. Later on, people took scores on IQ tests to represent the measure of an inherent (and immutable) capacity. That over-extension has been the cause of unending problems.

Along the fifth arrow: How well do the constructs and relationships in the conceptual model map back into the corresponding attributes of the original situation?

It must be stressed that constructs that seem important in the representational system may or may not have much explanatory power—or even be meaningful—in the conceptual model (or the situation from which the model was abstracted). This can easily occur when the constructs in the model are arrived at statistically—for example, when they are produced by methods such as factor analysis. "Verbal ability" in mathematical performance is one case in point.

As we complete the circuit in Figure 19.5, it is worth recalling that the last arrow represents the completion of the representation and analysis process—and that the process involves working with *selected and abstracted features* of the situations they represent. Any use of a model or representation idealizes and represents a subset of the objects and relationships of the situation being characterized. The conceptual model may cohere, and analyses within it may be clear and precise—but the whole process is no better than any of the mappings involved, especially the mapping back into the original situation. All results must be interpreted with due caution, for they reflect the assumptions made throughout the entire process.

One quantitative case in point was the use, in the 1960s and 1970s, of factor analyses to determine components of mathematical ability. Various tests were constructed to assess students' "verbal ability," "spatial ability," and more; then studies were done correlating such abilities with problem solving performance. Over time, however, it became clear that most such "abilities" were almost tautologically defined—that is, you had "verbal ability" to the degree that you scored well on tests of verbal ability. However, researchers were unable to explain how these abilities might actually contribute to competent performance.

If a whole field could delude itself in this way, imagine how easy it is for a single researcher to do the same. Much qualitative research consists of the construction of categories to represent perceived patterns of data. The analytic perspective that one brings to one's work may well shape what one sees or attends to—and thus which categories are constructed.

As an indication of the universe of possibilities, LeCompte and Preissle (1993) distill "major theoretical perspectives in the social sciences" into a table that covers six pages of small-sized print (pp.128–133). Those perspectives, accompanied by a few of their major theoretical constructs, are:

- Functionalism (systems, functions, goals, latent and manifest functions, adaptation integration, values, cultural rules…)
- Conflict theory (many of the same concepts as functionalism, plus in addition: legitimacy, consciousness, domination, coercion…)
- Symbolic interactionism and ethnomethodology (self, self-concept, mind, symbols, meaning, interaction, role, actor, role taking…)
- Critical theory (resistance, human agency, repression, hegemony, subjectivity, political economy, consciousness (false and true)…)
- Ethnoscience or cognitive anthropology (cultural knowledge, cognitive processes, cognitive models…)
- Exchange theory (cost, benefit, rationality, fair exchange, rewards, norms of reciprocity, satiation…)
- Psychodynamic theory (id, ego, superego, culture and personality, neurosis, psychosis…)
- Behaviorism (individual differences, stimulus, response, conditioning…)

Given this extraordinary diversity of perspectives and constructs, one must ask: how can the field sort out which ones make sense; which perspectives are relevant and appropriate to apply in which conditions; and, how much faith can one put in any perspective or claim? These questions are the focus of Section V.

V. STANDARDS FOR JUDGING THEORIES, MODELS, AND RESULTS[6]

Given the wide range of perspectives, methods and results in educational research, the following questions are essential to address. What grounds should be offered in favor of a general theory, or a model of a particular phenomenon? How much faith should one have in any particular result? What constitutes solid reason, what constitutes "evidence 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, generality, and trustworthiness
- Multiple sources of evidence (triangulation)

Each is briefly described in this section. In the next section, these criteria will be invoked when various types of research are considered.

Descriptive power

Descriptive power denotes the capacity of theories or models to capture "what counts" in ways that seem faithful to the phenomena being described. As Gaea Leinhardt (1998) 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 when one is exploring some of the cow's physiological properties.

Simply put: Theories of mind, problem solving, or teaching (for example) should include relevant and important aspects of thinking, problem solving, and teaching, respectively; they should capture things that "count" in reasonable ways. At a very broad level, it is fair to ask: Do the elements of the theory correspond to things that seem reasonable? And, is anything missing? For example, in the 1970s and 1980s researchers designed a fair number of data coding schemes, to "capture" the actions taken by people as they tried to solve mathematics problems (see Lucas, Branca, Goldberg, Kantowsky, Kellogg, & Smith, 1980, for one such example), or to capture classroom actions (see, e.g., Beeby, Burkhardt, & Fraser, 1979). Here is one test of its descriptive power. Suppose you study the coding scheme and become proficient at its use. Suppose further that someone else proficient in the use of the scheme makes a videotape of the relevant phenomenon, and then codes it according to the scheme. You are given the coding, which you examine. Then, when you look at the videotape, are there any "surprises"—relevant behaviors or actions that the coding scheme did not prepare you to see? If so, there is reason to question the descriptive adequacy of the scheme.

More broadly, there is the question of whether an analytic scheme or representation takes the right factors into account. Suppose someone analyzes a problem solving session, an interview, or a classroom lesson. Would another person who read the analysis and then saw the videotape, reasonably be surprised by things that were missing from the analysis? This might call into question the theoretical underpinnings of the approach. To take an historical example, consider the "process-product" approach, a once-dominant paradigm in studies of teaching (Brophy & Good, 1986). Researchers coded classroom behaviors (amount of time on task, frequency of direct questions asked of students, etc.) and then explored correlations between the extent of those behaviors and measures of student success, such as scores on standardized tests. Curiously absent from such studies (and easy to see in hindsight, though not at all apparent at the time) were what we now consider relevant cognitive considerations: What did it really mean to understand the mathematics? How was it explained? What content did the teacher and students focus on? How did the study of the relevant mathematical processes play out in the classroom, and how was it represented on the tests? With 20-20 hindsight we can see such omissions in methods of the recent past. We need to keep our eyes open for similar lapses in our current work.

Explanatory power

Explanatory power denotes the degree of explanation provided regarding 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. Consider, for example, the kinds of finely detailed coding schemes for problem solving behavior (Lucas et al., 1980) discussed above. They provided a wealth of detail regarding what the subjects *did* (along specific dimensions), but little relevant information regarding how and why the subjects were ultimately successful (or not) at solving the problems. Likewise for the process-product paradigm: "classroom processes" were hypothesized to be related to "learning" and "performance outcomes," but the mechanisms by which they were related went unexamined.

There are at present many alternative forms of explanation and descriptions of mechanism; the field will need to sort these out, over time. Cognitive explanations tend to focus on "what goes on in the head," at some level of detail. 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. A typical cognitive explanation would focus on a description of working memory. It would provide 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. Such explanations work at a level of mechanism: they say 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. Similarly, socio-culturally and anthropologically oriented research aimed at explaining what takes place in classrooms focuses on describing how and why things happen the way they do. There are, of course, myriad ways to do this. For example, Bauersfeld's (1980) article "Hidden dimensions in the so-called reality of the classroom" provides an alternative perspective on classroom events, elaborating on the "hidden agendas" of students and teachers. Boaler's (2002) study of reform and traditional instruction traces the impact of alternative classroom practices on students' performance and hypothesizes mechanisms to account for the very different patterns of gender-related performance in the two instructional contexts. Stigler and Hiebert (1999) provide coherent explanations for what might seem incidental or inexplicable phenomena. (For example, why is it that overhead projectors (OHPs), which are widely used in the United States, are rarely found in Japanese classrooms? After all, such technologies are easily accessible in Japan. The answer has to do with lesson coherence. OHPs are devices for focusing students' attention. As such, they fit in wonderfully with typical instructional patterns in the United States—they support teachers in saying "here is what you should be attending to, now!" A major goal of Japanese lessons, however, is to provide students with a coherent record of an unfolding story, reflecting the evolution of the lesson as a whole. Japanese teachers tend to make careful use of the entire white- or chalk-board, providing a cumulative record of how a whole lesson unfolds. The OHP, with its limited focus and ephemeral nature, is not suitable for this purpose.)

Scope

Scope denotes 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.

One reason that there is currently so much theoretical confusion is that adherents of one approach or another rarely delineate the set of phenomena to which their theories apply, and to which they do not. Buswell made this point a half-century ago:

> The very reason that there are conflicting theories of learning is that some theories seem to afford a better explanation of certain aspects or types of learning, while other theories stress the application of pertinent evidence or accepted principles to other aspects and types of learning. It should be remembered that the factual data on which all theories must be based are the same and equally accessible to all psychologists. Theories grow and are popularized because of their particular value in explaining the facts, but they are not always applied with equal emphasis to the whole range of facts. (Buswell, 1951, p. 144)

When he wrote, Buswell was explaining that behaviorism explained *some* things well (and still does) while not being of much use with regard to some other phenomena; likewise for "field theories" such as Gestaltism. But the point is general. And research will only make progress if researchers take care to specify what a theory (or a model, or piece of research) actually does, and does not.

One case in point is the Teacher Model Group's work studying teachers' in-the-moment decision-making in the classroom, called a "theory of teaching-in-context" (Schoenfeld, 1998, 1999c, 2002b). The goal of that research is to provide an explanation of every decision made by a teacher while engaged in the act of teaching, as a function of the teacher's knowledge, goals, and beliefs. This is ambitious, and the work is carried out at a very fine-grained level of detail. At the same time, the constraints of the theory and its associated models of teachers are carefully spelled out. It is *not* a theory of teaching (writ large), or a theory of "what happens in the classroom." For example, the theory provides a view of "classroom reality" only as seen from the teacher's point of view—each student's view will certainly differ, and that of an observer focusing on the class as a "dynamic entity" will differ as well. External constraints (e.g., the politics of schooling) are not modeled directly, although the teacher's perception of them is included as part of the model. Changes in the teacher (i.e., learning as a function of experience) are not modeled: the way the model works is that the teacher's understandings are modeled at the beginning of a lesson, and serve as the basis for the analysis that follows. That is, given what we know about the teacher (including his or her history with the students and understandings of them, understanding of content, etc.) *right now,* here is how he or she is likely to react to the "next" thing students do. In short, the research group has taken pains to specify what the theory of teaching-in-context does, and what it does not. It can then be held accountable (according to some of the criteria enunciated in this section) for the adequacy with which it addresses the phenomena it claims to address.

Predictive power

"Prediction" in education and the social sciences is a touchy business. Claiming to have a model of some form of behavior, or a model of an individual (e.g., modeling someone's teaching) is likely to raise hackles almost immediately. A typical reaction is, "People are individuals, they have free will, they make on-the-fly decisions; how can you possibly predict what they'll do?" And of course, one can't—in the sense of saying precisely how someone will act in any situation. The very idea of suggesting that one can predict someone's actions seems reductive and dehumanizing. Yet, prediction is possible and important, if not essential.

For those in the sciences, prediction is a *sine qua non* of theory. Most theories in mathematics and the sciences allow for predictions of the type, "In certain circumstances, when X happens, then Y happens." Of course, "Y happens" can take various forms. The kinds of predictions that make people nervous when they think about predictions of human behavior are those like the definitive predictions from classical mechanics (specifying the motion of particles subjected to specific forces) and chemistry (specifying the precise amount of radioactive decay, or the precise substances and quantities that will emerge from a chemical reaction). There are many other forms of prediction, however. Consider, for example, models of predator-prey relationships. Once the initial assumptions are fed into a model, the model predicts the change of the populations relative to each other. Such models predict very specific trends (with numbers attached), and the accuracy of the predictions can be measured against the actual populations of predators and prey. Predictions may be in the form of statistical distributions, as in the case of Mendelian genetics. In this case, evaluation of the predictions is easy: does the population of offspring have the distribution that the theory predicts? In other cases, predictions can be converted into statistical or probability distributions. Weather forecasting also gives rise to statistical distributions. The question is: over time, what per cent of the time did it rain, on those days when the forecaster said there was a (say) 30% chance of rain? Also, predictions may be in the form of *constraints*—statements of what is possible or impossible. Evolutionary theory is a case in point. Whatever evolutionary theory is proposed must apply not only to known data but to previously unexamined fossil records as well. That is, the theory predicts what properties sequences of fossils in geological strata can or cannot have. A cumulative fossil record consistent with the theory is taken as substantiation for the theory—and any discrepant fossil record that is discovered will be considered *very* problematic for it. In short, even theories such as evolution, which are anything but deterministic, support strong predictions. The question for educational studies is, what kinds of predictions does a proposed theory support?

Sometimes it is possible to make precise predictions. For example, Brown and Burton (1978) 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 ("bugs"), 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 specific incorrect answer that a 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, or a model of a particular teacher, can make specific predictions of the kinds just discussed. It can suggest constraints ("in these circumstances, this teacher will *not* do the following…"), and it can suggest likely events ("Given this chain of events, there is a 70% chance the teacher will respond in the following way, and a 30% chance the teacher will respond this way instead"). Such predictions can be made without being either reductive or dehumanizing.

It should also be noted that 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 serious 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? Of course, 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 physical objects are. 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.[7] 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 some quick examples. The model of the teaching process constructed by the Teacher Model Group (Schoenfeld, 1998, 1999c, 2000a, 2002a, in press) includes components that represent aspects of the teacher's knowledge, goals, beliefs, and decision-making.

Skeptics (including the authors) should ask, how clear is the representation? Once terms are defined in the model (i.e., once a teacher's knowledge, goals, and beliefs are described) is there hand-waving when claims are made regarding 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? These criteria—are the objects and relations in the model well specified, and is the correspondence between those entities and the entities they are supposed to represent clearly delineated? —should be applied whenever researchers claim to have a model of some phenomenon. For example, Lesh and Kelley (2000) claim that there are three levels of models in their three-tiered teaching experiments—models created by students, teachers, and researchers. Are the models specified and inspectable? Similarly, "APOS theory" (see Asiala, Brown, de Vries, Dubinsky, Mathews, & Thomas, 1996) uses terms such as Action, Process, Object, and Schema. Would you know one if you met one? Are they well defined? Are the ways in which they interact or become transformed well specified? In all these cases, the bottom line issues are, "What are the odds that the so-called theory or model is a phlogiston-like theory or model? 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." They should be applied to all of the theoretical constructs in the long list that ended Section IV. (In the biased view of this author, many if not most of the constructs fail the test. We have our work cut out for us.)

Falsifiability

The need for falsifiability—for making non-tautological claims or predictions whose accuracy can be tested empirically—should be clear at this point. Simply put: if you can't be proven wrong, you don't have a theory.[8,9] A field makes progress (and guards against tautologies) by putting its ideas on the line.

Replicability, generality, and trustworthiness

Replicability, like prediction, is controversial. It should be, if one takes the spirit and meaning of replicability from the experimental sciences: if one does "exactly the same thing," will the same results occur? Given the variability of people and contexts, that strict notion of replicability is rarely appropriate for educational research. Moreover, one should not expect many educational studies to be replicable—there is a wide range of studies that deepen our understandings without making general claims. For example, biographical studies may help readers understand how certain forces shaped the lives of certain individuals, without claiming that others would necessarily act in the same way. Studies of how attempts at "reform" played out in various school districts are similarly not replicable: readers may derive important lessons from them, but there is no expectation that similar attempts at reform in similar school districts will necessarily play out in the same ways. Likewise, some studies of teaching may have the primary value of enhancing readers' understandings of the subtleties and complexities of teachers' classroom actions and what drive them. Cooney's (1985) study of "Fred" shows how a teacher can avow the importance of teaching for problem solving but come to teach in a very traditional way. Cooney presents evidence that Fred, despite his rhetorical homage to Pólya, understood problem solving to be a motivational device rather than a way of engaging in mathematics. Hence, when students did not value his use of motivational problems, *and* he felt pressure to make sure that the students understood core content, he jettisoned "problem solving" to spend more time on "basics." David Cohen's (1990) study of "Mrs. Oublier" provides similar insights. Mrs. Oublier claimed to have adopted reform methods—but her understanding of reform was rather shallow, and many of her established teaching habits undermined her attempts to adopt reform practices. From such studies readers learn to look at teaching in more subtle, nuanced ways—but they do not expect other teachers to behave precisely the ways that Fred or Mrs. Oublier did.

Replicability is an issue, however, when theoretical claims are made; also when claims are made regarding the generality of various phenomena. If a theory posits that people have certain mental structures, for example, then other researchers should expect to document the existence of such structures. A paradigmatic case is that of short-term memory (STM). George Miller's famous 1956 paper "The Magic Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information" makes the claim that the capacity of short-term memory is strictly limited—that people typically have between five and nine short term memory "buffers" that hold information, temporarily, while performing mental actions. Such a limitation would put serious constraints on the capacity of individuals to perform a wide range of mental actions. For example, multiplying two three-digit numbers, say 384 × 673, requires keeping track of more than 9 subtotals. Most people will not be able to perform this task with their eyes closed, because they will forget some of the numbers involved before they can complete the task. This finding can be easily replicated, and the fact that it can be establishes the robustness of the finding.[10,11] Claims about other cognitive structures or patterns can be subjected to comparable tests of robustness. For example, much of the early work on aspects of metacognition, or on the development and impact of beliefs on students' mathematical performance has been replicated, with students at various age levels.

Similar observations can be made with regard to sociocultural or ethnographic perspectives. Of course, many such studies do not bear replication: they provide insights into particular situations and contexts, which cannot be "duplicated" in any meaningful sense. However, one can examine the robustness of theoretical constructs, by asking about the consistency with which they are applicable and informative in contexts where they are said to apply. For example, the idea of the "didactical contract" (see, e.g., Brousseau, 1997) has been at the foundation of a large body of French educational research for some decades, and has provided a consistent and productive orientation to empirical studies.

It should be noted that issues of replicability, generality, and trustworthiness are deeply connected to the issues of rigor and specificity discussed above. The ability to replicate a study, or to employ a theoretical construct in the way it was employed by an author, depends on the original work being well enough defined that other researchers following in the footsteps of the authors can employ methods or perspectives that are quite close to the original. This should be obvious, but it has not been, historically. Consider this case in point from the classical education literature. Ausubel's (1968) theory of "advance organizers" 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 more than a decade 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.

There are, of course, standard techniques within both the cognitive and anthropological traditions for dealing with such issues. One is the training of multiple observers. A series of trained observer-analysts can work through the same body of data independently, and compare notes afterwards. If all goes well, those observers should "see" pretty much the same things. And, if the constructs are truly well defined and communicated, researchers from outside the original research group should to be able to learn the techniques and, when given the data (such as videotapes, etc) draw essentially the same conclusions. It should be noted that the cognitive/experimental and social/anthropological communities each have their own approaches to these issues, but that there is overlap in spirit if not in detail. Within the cognitive community, for example, there is a tradition of computing inter-rater reliability to identify the degree to which independent researchers assign the same coding to a body of data (say a transcript or videotape). Those who work within anthropological traditions tend to discuss the "trustworthiness" of a study. For a discussion of the relationship between these two traditions, see Moschkovich and Brenner (2000). For more extended discussions of these constructs, see LeCompte, Millroy, and Preissle (1992), LeCompte and Preissle (1993), and Lincoln and Guba (1985).

One source of trustworthiness is having multiple eyes view the same data. Another, to which we now turn, is having multiple lines of evidence or argument that point to the same interpretations or conclusions.

Multiple sources of evidence ("triangulation")

Argumentation in education is much more complex than in mathematics and the physical sciences. In mathematics, one compelling line of argument (a proof) is enough: validity is established. In education (more broadly, in the social sciences), we are generally in the business of looking for *compelling evidence*. The fact is, evidence can be misleading—what one thinks 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 they were 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, with 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 picture. In modeling teaching, for example, the Teacher Model Group draws inferences about the teacher's behavior from videotapes of the teacher in action—but it also conducts interviews with the teacher, reviews his or her lesson plans and class notes, and discusses tentative findings with the teacher. In this way, the group deliberately seeks convergence of the data. The more independent sources of confirmation there are, the more robust a finding is likely to be.

For additional discussions of the issues discussed in this section of this chapter, see Clement (2000), Cobb (2000), and Moschkovich and Brenner (2000). Clement's comments are grounded in his experience using clinical interviews to build models of students' understandings of a range of science concepts. A key concept for Clement is the *viability* of a model. He offers (p. 560) a set of criteria for evaluating the viability of models and theories that overlaps significantly with those discussed here. Cobb, in a discussion grounded in his experience with teaching experiments, focuses on the *generalizability* and *trustworthiness* of analyses. Moschkovich and Brenner provide an overview of both traditional and naturalistic approaches to these issues.

VI. A HEURISTIC FRAMEWORK FOR SITUATING RESEARCH STUDIES, AND A SET OF ISSUES IT RAISES

Prologue: A structural dilemma

Researchers in (mathematics) education now have access to an extraordinarily wide array of methods. They confront enduring questions regarding which kinds of methods are appropriate in which circumstances, a problem exacerbated by the variety of methods currently available. My original intention for this section of this chapter was to provide a selective overview of some relevant categories of research methods, and to raise some issues about their use. This is by no means a straightforward task. Indeed, as I worked to organize this section, I came to realize that the very notion of an "overview of methods" is likely to be a fruitless endeavor. More central, and more to the point, are questions regarding the purposes of the research undertaken and the kinds of information that various research methods can yield—when those are understood, the selection of methods and their application should follow. Thus, rather than offering a taxonomy of methods, this section of this chapter will offer what might be considered a heuristic guide to thinking about different kinds of claims that are made in educational research, and the warrants researchers might produce to justify those claims.

Since the idea of a taxonomy of methods has clear face validity and might seem natural to the reader, it is worth explaining why that approach was abandoned. When I constructed the outline of this chapter, it seemed logical that at some point I would discuss what might be considered "rough equivalence classes" of approaches to research, raising some issues concerning the character of each. If one decides to take that approach, it is hardly necessary to reinvent the wheel; others have produced state-of-the-art categorizations. It seemed reasonable, therefore, to base the taxonomy on recent categorizations of current research. An obvious candidate for a starting point was the *Handbook of Research Design in Mathematics and Science Education* (Kelley & Lesh, 2000). Its editors chose to emphasize research designs that are intended to radically increase the relevance of research to practice. Examples of such research designs include:

- Teaching experiments
- Clinical interviews
- Analyses of videotapes
- Action research studies
- Ethnographic observations
- Software development studies
- Computer modeling studies (Kelley & Lesh, 2000, p. 18)

My expectation was that I would supplement this categorization of designs (many of which, like design experiments, reside in "Pasteur's quadrant") with discussions of more traditional approaches to educational research, such as experimental designs and statistical studies.

Such an approach turned out to be impossible. The reason is that on closer examination the set of categories given above turns out to be fundamentally incoherent. This incoherence is on at least two dimensions: ill-definedness and categorical overlap. Regarding the former, consider, for example, the term *teaching experiment*.

> In general, teaching experiments focus on development that occurs within conceptually rich environments that are explicitly designed to optimize the chances that relevant developments will occur in forms that can be observed. The time periods that are involved may range from a few hours, to a week, to a semester or an academic year. Furthermore the environment being observed may range from small laboratory-like interview rooms, to full classrooms, to even larger learning environments. (Kelley & Lesh, p. 192)

Such an all-encompassing definition allows for studies that bear little resemblance to each other—with regard to any of context, focus, or investigatory method(s) —to be considered members of the same category. In the *Handbook's* section on teaching experiments, for example, Lesh & Kelley (2000) describe "multitiered" teaching experiments in which teams of students "work on a series of model-eliciting activities," participating teachers "construct and refine models to make sense of students' modeling activities, and researchers "develop models to make sense of teachers' and students' modeling activities." In a chapter, entitled "teaching experiment methodology: Underlying principles and essential elements," Steffe and Thompson (2000) focus on developing an understanding of "students' mathematics"—"whatever might constitute students' mathematical realities" (p. 268). Their goals are in many ways consonant with the goals of traditional experimental studies, although their methods are radically different. Their teaching experiment was, in essence, a form of hypothesis testing as follows: "Suppose we identify two groups of students who (as far as we can tell) have developed for themselves different understandings of the counting process.[12] We hypothesize that these two groups of students will respond differentially to a particular kind of instruction, with the gaps between the two groups increasing as a result of instruction." In short, Steffe and Thompson were conducting an experiment with the expectation that students' specific (attributed) cognitive structures would interact with instruction in particular ways. They note the following:

> We use *experiment* in "teaching experiment" in a scientific sense. The hypotheses in the teaching experiment [described immediately above] were that the differences between children of different groups would become quite large over the 2-year period and that the children within a group would remain essentially alike. That the hypotheses were confirmed is important, but only incidental to our purposes here. *What is important is that teaching experiments are done to test hypotheses as well as to generate them. One does not embark on the intensive work of a teaching experiment without having major research hypotheses to test.* (Steffe & Thompson, 2000, p. 277; emphasis added)

There are, indeed, some similarities between the studies reported by Lesh and Kelley and by Steffe and Thompson, including the iterative and reflective character of the studies. But the differences far outweigh the similarities. Lesh and Kelley characterize their work as "a longitudinal development study in a conceptually rich environment" (page 197), while Steffe and Thompson characterize their work as a (very rich) form of hypothesis-testing experiment.

These two examples alone point to the fundamental incoherence of the category—and the problem gets worse when one considers other examples of the category given in the *Handbook,* or classical examples such as those found in the *Soviet Studies in School Mathematics* (Kilpatrick & Wirszup, 1975).

This problem was exacerbated by the significant overlap of the various categories listed above. For example "analyses of videotapes" are employed in all of the categories listed above—in teaching experiments, clinical interviews, action research, ethnographies, and computer-based development and modeling studies. A large number of "action research studies" are teaching experiments (in the sense defined by Lesh and Kelley, above) and vice versa. Software development studies often involve teaching experiments as part of their design. And so on.

In short, the kind of taxonomy offered by the *Handbook*—the kind of taxonomy I had hoped to use as the basis for this section of this chapter—is fundamentally incoherent. It is based on surface structure rather than deep structure. The problem for this chapter, then—and for the field—becomes, *What is an appropriate deep structure for conceptualizing and organizing research in (mathematics) education?*

Would that there were a straightforward or clear answer to this question. This section offers one tentative approach, which can be considered preliminary at best. In keeping with much of the qualitative literature (see, e.g., Cobb, 2000), I shall argue that, whether one is discussing quantitative or qualitative research, *generality* (or scope) and *trustworthiness* are two fundamental dimensions of research findings; and that *importance* is a third. In what follows, I shall briefly elaborate on this perspective. Having done so, I shall use this frame to structure the discussion of a number of illustrative examples. The goal is to provide a way of thinking about the implications of various findings—how well they are warranted, and how widely they apply. The discussion will proceed along the "generality" dimension of the framework. I start with examples of little generality, and discuss their properties (specifically, their trustworthiness and importance). The discussion then proceeds through a series of examples of increasing generality.

A provisional organizational frame

Figure 19.6 provides a schematic representation of a three-dimensional framework for considering the character of research studies in education. As suggested immediately above, three main attributes by which a study can be judged are the following:

- *Generality, or Scope*. The *claimed generality* of a study is the set of circumstances in which the author(s) of a study claim that the findings of the study apply. The potential generality of a study is the set of circumstances in which the results of the study (if trustworthy) might reasonably be expected to apply.
- *Trustworthiness*. The issue is, how well substantiated (according to many of the criteria elaborated in Section V) is the claimed generality of the study? How solid are the warrants for the claims? Do they truly apply in the circumstances in which the authors assert that the results hold?
- *Importance:* To put things bluntly, how much should readers care about the results?

Here are a few examples to clarify the way these constructs will be used.

One classic study, conducted by Harold Fawcett, is reported in the 1938 *Yearbook* of the U. S. National Council of Teachers of Mathematics, *The Nature of Proof.* Fawcett provides a richly textured description of a course in geometry that he developed and taught. The fundamental goals of the course were to (a) help students develop a deep understanding of the concept of proof in mathematics, through the study of geometry; and (b) to link those

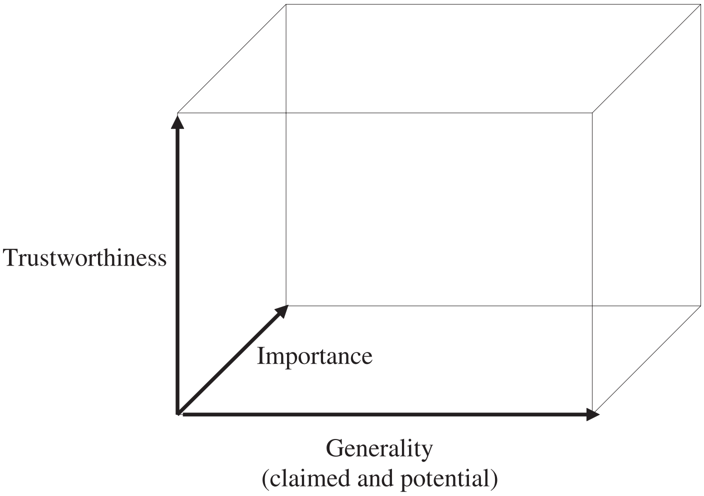


Figure 19.6 Three important dimensions along which studies can be characterized.

understandings, and the reasoning processes involved, to real world deductive reasoning. Fawcett describes the nature of instruction with some care. Readers get a sense of classroom discourse, of the kinds of questions the class debated, and of the flow of argument. Fawcett provides instructional artifacts, such as the forms students used to analyze arguments in the media, and he describes classroom discussions concerning those arguments. He provides a list of geometric results derived by the class, so readers can develop a good sense of the curriculum. And, he offers multiple forms of evaluations of student performance in the course: student scores on a state-wide test of plane geometry; a "transfer" test of reasoning in non-mathematical situations; data from students regarding their reasoning outside the course; comments from parents regarding their children's abilities to think critically; comments from six external observers; and student testimonials.

The body of evidence offered by Fawcett is compelling and convincing—hence his research report scores quite high on the *trustworthiness* dimension. Its score on the *generality* dimension, in contrast, is quite low; Fawcett has offered compelling detail in the description of one class, and a rather unusual one at that. His report is, in essence, an existence proof. Fawcett has shown that it is possible to offer instruction from which students can develop a deep understanding of geometry, and that his students were able to apply the reasoning skills that they learned in the course to "every day" arguments as well. From my perspective, that makes his findings quite important—it shows that such goals can be achieved. (Consider by analogy another existence proof, Orville and Wilbur Wright's flights at Kitty Hawk in 1903. The Wright brothers made four flights in one day, the last of which lasted 59 seconds and covered 852 feet. The evidence of engine-powered heavier-than-air flight was trustworthy. There was no generalization at that point, but the very fact that flight was achieved ultimately opened up a world of possibilities.) As will be seen below, a fair number of important educational studies are of this type. In addition, pioneering studies are often followed by replications and extensions, which serve to establish the generality of the findings.

As a second example, in the 1970s and 1980s there were a large number of studies (e.g., Clement, 1982; Clement, Lochhead, & Monk, 1981; Rosnick & Clement, 1980) that dealt with the "students and professors" problem:

> Using the letter S to represent the number of students at this university and the letter P to represent the number professors, write an equation that summarizes the following sentence: "There are six times as many students as professors at this university."

Numerous replications of the original studies in a wide range of contexts indicated that more than half of the undergraduate students not majoring in mathematics, science, or engineering produce the equation $P = 6S$ instead of $S = 6P$. A wide variety of explanations for this phenomenon were offered, and some attempts at remediation were made on the basis of those explanations. After some time, however, the well ran dry. Compelling explanations of the phenomenon did not emerge, and attempts at remediation were not demonstrably successful. The field's attention turned elsewhere.

In terms of the criteria discussed above, this body of research is trustworthy—at least in the sense that the phenomenon was robust and easily replicated. Generality is relatively low, in that the phenomenon, while robust, was never tied to any theoretical ideas that had significant scope. And ultimately, the findings—though "hot" for some time—were relatively unimportant.

Note that there can be very different sources of trustworthiness, depending on the nature of the claims and the methods involved. Fawcett's work, though "small *n*," was trustworthy because of the richness of the analyses and consistency of the data. The "students and professors" data were trustworthy because of the replicability of the phenomenon and the large number of data points involved. Note also that large *n* in the latter example did not imply generality. Indeed, significant generality may be *suggested* by small *n* studies. For example, early studies of monitoring and self-regulation in mathematical problem solving suggested that the absence of effective metacognitive skills could be a cause of problem solving failure in *any* domain. And, large *n* is no guarantee of either trustworthiness or generality. For example, early studies regarding the effects of caffeine consumption turned out to be invalid because researchers neglected to take into account the correlation between coffee drinking and smoking, and were thus unknowingly conflating the effects of smoking and caffeine consumption. And, the findings of decades' worth of medical studies conducted with solely male samples were *assumed*—incorrectly, it turns out—to apply to females as well. The studies were far less general than originally thought. (Recall Figure 19.5: the choices of conceptual model and of focal variables, whether consciously or unconsciously made, have a fundamental impact on both the generality and trustworthiness of a study.)

The discussion that follows examines a series of research studies, ordered roughly by the generality of the claims made for them. For each study, methodological issues related to trustworthiness are discussed—the question being, "for studies of this type, what does it take to provide adequate warrants for the claims being made?"

Two points should be made regarding the choice of generality as the dimension along which studies will be ordered. First, this approach explicitly renders irrelevant the "qualitative/quantitative" distinction that has bedeviled the literature. The issues that count are the following: What kinds of claims are being made? What methods are appropriate for making those claims? What warrants are offered in defense of those claims? Providing trustworthy documentation of any particular kind of claim may call for quantitative methods; it may call for qualitative methods; it may call for both.

Another possible bifurcation is to separate research studies into the following two classes: research that tries to describe "things as they are," and research that documents attempts at change.[13] "Descriptions of things as they are" consist of attempts to describe objects, events, structures, and relationships as they occur. One obvious set of such studies consists of "naturalistic" observations. However, this class of studies is much broader: various probes, experimental or otherwise, are often used in order to discover what things are and how they work. For example, Piaget's claim that "object permanence" is learned rather than innate was established through a series of clinical interventions with young children. (Piaget obscured an infant's view of a key ring just as the child was in the middle of reaching for it, and the child stopped in mid-reach.) The same holds for almost all Piagetian clinical interviews, such as those regarding conservation of volume, the child's sense of time and space, and more. Piaget's goal was to develop an understanding of underlying cognitive structures and their development. He did so by confronting his interview subjects with interesting (and very carefully designed) situations, and then drawing inferences about the interviewees' underlying cognitive structures from their responses to the situations. Similarly, laboratory studies aimed at determining how many "buffers" people have in short-term memory are attempts to describe stable cognitive structures. And, large-scale testing often aims at descriptions of how things are. One example is Artigue's statement that "More 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" (p. 1379). In sum, the category "descriptions of how things are" is quite broad, and the methods used extraordinarily diverse.

On the surface, descriptions of action research or "attempts to make change and document it" might seem to be different from the type of study described immediately above. Much such work (e.g., Fawcett's, my problem solving courses, or various "design experiments") consists of attempts to establish existence proofs—attempts to show that something *can* be done. Other studies are comparative: the claims regarding the implementation of various kinds of software, or other instructional practices, are that students do "better" under certain conditions than under others. Yet, the methods used to document the claims often overlap with those used for descriptions of things-as-they-are. More importantly, the underlying issues concern questions such as: What kinds of claims are being made, and why should one believe them?" Like those mentioned above, these claims can be ordered by generality; then, given the nature of the claims, one can examine their trustworthiness. For these reasons, descriptions of things-as-they-are and things-as-they-might-be will be conjoined in the discussion that follows.

A spectrum of studies, ordered by generality

Category 1: Limited generality, but … (if properly done) …
"Here is something worth paying attention to."

A large number of studies are important not because they provide documentation of phenomena that are widespread, but because they bring readers' attention to an issue worth considering. The studies themselves may have very limited generality, but they may have heuristic value—they may point to issues that are important to consider, and may turn out to be general. They may deepen an understanding of some phenomenon. They may make a methodological contribution, or they may clarify or expand a theory.

For example, various studies have suggested that the "lessons learned" in classrooms can be very different than intended. One case in point is Wertheimer's (1945) argument, quoted in Section II, that instruction that focuses on drill and practice "is dangerous because it easily induces habits of sheer mechanized action, blindness, tendencies to perform slavishly instead of thinking, instead of facing a problem freely." On the one hand, the reader may well resonate with Wertheimer's claim on the basis of personal experience or classroom observations. On the other hand, one has to recognize that by contemporary standards, the evidence he offers in support of his claim is no more than anecdotal. The observations are not fleshed out in detail. One knows little about the background and classroom experiences of the students. There is little sense of how prevalent the phenomenon might be, or of how deep and resistant to change it is. (Were the results perhaps artifacts of his interactions with the students? Might the students have acted differently if he had structured the conversations somewhat differently?) The point here is not to chastise Wertheimer—those were different times, with different standards—but to point out that his observations, no matter how intriguing and important (and they were!), were not *trustworthy* in the sense of meeting the criteria elaborated in Section V. That trustworthiness can be compared with, say the descriptions of "making sense of linear functions" and "Hawaiian children's understanding of money" found in Moschkovich and Brenner (2000). In those studies (explicitly chosen as cases illustrating the integration of "a naturalistic paradigm" into research on mathematical cognition and learning), the authors explicitly address the questions one would expect the skeptical reader to pose: How credible are the claims? How broadly might they apply, and why should one believe that they do? How rich are the descriptions of events? What kinds of sampling was done? What kinds of triangulation? Did the researchers create an "audit trail" and make it accessible? When the answers to such questions are available and inspectable, readers can assess the degree to which the findings are trustworthy.

A very large percentage of educational studies are of the type, "here is a perspective, phenomenon, or interpretation worth attending to." The ultimate value of such papers is twofold, in that they offer a heuristic perspective ("one should pay attention to this aspect of reality") and because they can serve as catalysts for further investigation. As a case in point, consider Bauersfeld's 1980 paper, "Hidden Dimensions in the So-called Reality of a Mathematics Classroom." Bauersfeld re-interprets a teaching episode that had been the subject of another scholar's analysis. His "text" was a dissertation by G. B. Shirk (1972) at the University of Illinois, in which Shirk focused largely on the content and pedagogical goals of beginning teachers. Bauersfeld wanted to highlight a metatheoretical point—that teaching is a social activity as well as a cognitive one, and that viewing teaching as such can yield powerful insights into what happens in classrooms. His re-analysis "is used to identify four hidden dimensions in the classroom process and thus deficient areas of research: the constitution of meaning through human interaction, the impact of institutional settings, the development of personality, and the process of reducing classroom complexity" (Bauersfeld, 1980, p. 109). The phenomena were not (yet) claimed to be general, but were declared to be worthy of investigation. Similarly, various studies of discourse in classrooms, illustrating analyses from a "situative perspective" (see, e.g., Greeno & the Middle-School Mathematics through Applications Project Group, 1997, 1998), serve the joint purpose of illuminating a set of particular classroom events and highlighting the potential value of an emerging theoretical approach.

Many other studies do not make such claims overtly, but in essence have similar intentions. Consider three teaching studies, which are in some ways similar and in some ways very different. Cooney's (1985) study of a beginning teacher showed how a beginning teacher's professed instructional goals could be undermined by his deeply held beliefs and his interactions with students. Cohen's (1990) study of a teacher undertaking "reform" showed how the teacher's well-established classroom routines could result in the perception but not the substance of reform:

> In the mid 1980s, California state officials launched an ambitious effort to revise mathematics teaching and learning. The aim was to replace mechanical memorization with mathematical understanding. This essay considers one teacher's response… she sees herself as a success for the policy: she believes that she has revolutionized her mathematics teaching. But observations of her classroom reveal that the innovations in her teaching have been filtered through a very traditional approach to instruction. The result is a remarkable melange of novel and traditional material. Policy has affected practice in this case, but practice has had an even greater effect on policy. (Cohen, 1990, p. 311)

In a third, richly detailed study, Eisenhart et al. portray the myriad factors that shape a student teacher's decision making:

> We reveal a pattern in which … there were a variety of strong commitments to teaching for both procedural and conceptual knowledge; but … the student teacher taught, learned to teach, and had opportunities to learn to teach for procedural knowledge more often than and more consistently than she did for conceptual knowledge. We find that the actual teaching pattern (what was done) was the product of unresolved tensions within the student teacher, the other key actors in her environment, and the learning-to-teach environment itself. (Eisenhart et al, 1993, p. 8)

In all of these studies, there are suggestions of generalizable findings: teacher goals can be subverted if they are not tied to meaningful, implementable ideas (Cooney); some instructional goals are sufficiently nebulous that teachers can believe they are attaining them when they are not (Cohen); and, conflicting pressures and mixed messages from the school district and state, along with shaky subject matter knowledge, can undermine the intention to teach for concepts as well as skills (Eisenhart et al., 1993). The suggested generality of these findings, and the fact that attempts at teaching for understanding might be undermined if they are not taken into account, is what makes them *important*. The implications are heuristic, however: "we believe there are many cases like this in the world, and it would be good to keep the implications of these studies in mind." The claims themselves are not about generality: the evidence offered is about the cases at hand. The standard for judging these papers, given their claims, is: are the cases compelling, and the analyses trustworthy? Making that decision entails, of course, judging whether the methods employed provide adequate evidence for claims made (and whether they provide evidence or arguments that counter alternative explanations).

As noted above, various existence proofs also fall into this category. Fawcett's (1938) study demonstrated that students could, under appropriate circumstances, learn aspects of formal mathematical arguments and apply their understandings in real-world contexts. The same is the case for various design experiments (e.g., Brown, 1992; Brown & Campione, 1996) and fine-grained research on various other instructional interventions (see, e.g., Cognition and Technology Group at Vanderbilt, 1997; Schauble & Glaser, 1996).

A word about "design experiments" is in order here. The term was invented in order to justify the idea that scientific work could be done in the context of real-world interventions, and to offer an alternative to the standard model of experimentation, where "treatments" and outcome measures are designed in advance. The underlying idea is that a complex intervention is planned and implemented, and huge amounts of data (including videotapes, class logs, student work, etc.) are gathered. If interesting or important events appear to take place, the data are analyzed (depending on the nature of the events) to document their existence or explain their occurrence. Some of these explanations are post hoc: the events are noted, and the record is combed for relevant evidence. But the order of data gathering is not essential. What is essential is the following: once claims are made, how do they stack up against the criteria identified in Section V? Do the methods employed provide some substantial degree of trustworthiness regarding the findings? As such, the methodological issues concerning such studies are similar in kind to those concerning other studies described in this category. (If broader claims are made regarding design experiments or other interventions, then the studies fall into the next category.) For extended discussions of design experiments, which are still an unsettled construct, see Kelley (2003) or Schoenfeld (2006).

Category 2: Some generality is claimed

One case in point here is the quotation from Artigue (1999) given above: "More 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" (p. 1379). Similar kinds of statements are made regarding various national assessments of mathematical competency, cross-national comparisons, and so on. When such statistical statements are made, there are issues of sampling, of construct validity (does the question warrant the interpretation given?), and more—recall Figure 19.5.

Other statements concerning the typicality of various phenomena—especially phenomena not amenable to testing of the type just described—may come with different kinds of warrants. Here are two examples that claim some degree of generality, but do not quantify it.

In her studies of mathematics teachers' knowledge, Liping Ma (1999) analyzed Chinese and U.S. teachers' responses to a series of questions regarding topics or problems in elementary mathematics and how they might teach them. Ma's sample of teachers included 23 "above average" teachers from the United States, and 72 teachers of a wide range of ability from China. Her research documents, with care and detail, the fact that the sample of Chinese teachers had a deeper knowledge of mathematics and how to teach it than did their U.S. counterparts. Specifically, eight of the Chinese teachers had developed a form of understanding that Ma calls "profound understanding of fundamental mathematics"—a rich and connected view of the content and ways to promote student learning of it. None of the U.S. teachers had developed comparable knowledge.

Ma does not focus on the statistics. Hers were not random samples, and there is no claim that her statistics represent the percentages of the populations of U.S. and Chinese teachers who have developed a profound understanding of fundamental mathematics. Nonetheless, the differences in percentages are dramatic. They suggest strongly that a non-trivial percentage of teachers in China develop this deep form of knowledge, and that it is relatively rare among teachers in the United States. Indeed, the way that Ma's samples were constructed lends additional credence to those findings: her sample included a spectrum of Chinese teachers, while the teachers from the United States were considered "above average." Hence, in addition to the trustworthiness of her analysis, there is a plausible degree of generality to her findings. The richness of the analysis lends plausibility to the generality of the findings, even though no claim is made for it.

A similar suggestion of generality, without precise statistics, could be seen in a series of studies I conducted regarding student beliefs about learning and doing mathematics. In a series of observations in one focal classroom school, I documented instructional practices, including the fact that a typical test contained 25 problems to be worked in 54 minutes, and that in a typical class period, students would work more than a dozen problems. The documentation included statements from the teacher to the effect that students would not have time to think through problem solutions on tests; they would have to enter the exam knowing how to solve the problems they would face. Students were interviewed, and they were videotaped as they worked on problems. I also administered a questionnaire to 230 students (including the focal class) at various grade levels in the metropolitan area containing the school. Among the data gathered were the following.

> The 206 responses to the question "how long should it take to solve a typical homework problem" averaged just under 2 minutes, and not a single response allotted more than five minutes. The largest of the 215 responses to "What is a reasonable amount of time to work on a problem before you know it's impossible?" was twenty minutes; the average was twelve minutes. The following responses to both questions were typical. "Up to 2 or 3 minutes. I would work on a problem for about 10 minutes before deciding it's impossible." "A typical homework problem would take about 45 seconds. About 10 minutes for the impossible problem." "It would probably take from 30 seconds to 2 minutes. I usually give up after 3 or 4 minutes if I can't do it." "It should only take a few minutes if you understand it. No more than 10-15 minutes should be spent on a problem. (Schoenfeld, 1989, p. 340)

This kind of analysis led to the following conclusions.

> The data from this study help to provide a link between the fine-grained but small-scale observations in the [focal] study and the coarse-grained but nationwide data gathered in surveys such as the [U. S.] National Assessment. The questionnaire was administered in highly regarded schools with good graduation and placement rates. … The rhetoric of problem solving … was frequently heard in the classes we observed – but the reality of those classrooms is that real problems were few and far between, if they were seen at all. Virtually all of the problems the students were asked to solve were bite-size exercises designed to achieve subject matter mastery; the exceptions were clearly peripheral tasks that the students found enjoyable, but that they considered to be recreations or rewards rather than the substance of what they were expected to learn. This kind of experience, year after year, has predictable consequences. Students come to expect typical homework and test problems to yield to their efforts in a minute or two, and most of them come to believe that any problem that fails to yield to their efforts in twelve minutes of work will turn out to be impossible. (Schoenfeld, 1989, p. 348)

At issue here is the set of warrants for generality. Fine-grained studies of a focal class provided a possible explanation of mechanism, and a description of classroom practices allowed readers to decide whether these practices seemed typical. Statistical analyses of the focal classroom revealed no differences between their responses to the questionnaire and those of the larger group of 230 students, from a number of schools (which used state-wide curricula). And, the student responses on questions that overlapped with national assessments suggested that their responses were typical of responses nation-wide. This web of connections at least lends credence to the claim that the pattern of activities seen in the focal classroom, and their consequences, were anything but anomalous.

Other such broad notions of (typically unquantified) generality can be seen in research on aspects of thinking and learning such as metacognition. The general claim, broadly substantiated in the literature, is that the absence of effective self-monitoring and self-regulatory behavior is a significant cause of student failure in problem solving. Understanding this statement depends among other things, on one's definition of "problem solving." The circumstances in which it applies are those in which the problem solver is confronted with a task for which there is no obvious solution path, and decisions about how to approach the problem must be made. The claim is unlikely to apply to any significant degree in contexts in which problem solvers know the relevant techniques.[14] The methodological point here is that the "operating conditions" for many general claims need to be specified. Saying "X is important" implies across-the-boards generality, and appropriate warrants should be produced. Saying "X is important, and is likely to manifest itself in these particular circumstances" calls for a different set of warrants.

One final example of not-quite-specified-but-important generality deals with claims about attributes of particular groups. A paradigmatic example is the claim in Stigler and Hiebert (1999) that "teaching is a cultural activity." As a generalization, this kind of statement is of important heuristic value—and the authors make a good case for it. The warrant is that along certain dimensions, there is much more across-nation variation than there is within-nation variation. The devil is in the details—which in this case concern dimensions such as lesson coherence, time spent on individual exercises or problems, underlying conceptions of subject matter understanding, the use of instructional artifacts, and so on. The question for readers is, how solid are the warrants along the particular dimensions identified, and how well do differences in performance along those dimensions justify the general claims? (Here as with all of the other studies discussed, the criteria discussed in Section V can be applied to claims and the warrants provided for them.)

Category 3: Significant generality, if not universality, is claimed

Some years ago, Henry Pollak, in discussing differences between research in mathematics education and in mathematics, said, "there are no theorems in mathematics education." By that he meant that there are no abstract proofs that something *must* be the case; instead, evidence is offered until the conclusion seems established to the legal criterion—"beyond a reasonable doubt."

The fact is that certain claims in education are universals—typically, claims about underlying cognitive mechanisms or structures. Here are two familiar examples, mentioned earlier in this chapter.

As noted above, Piaget documented the fact that children are not born with "object permanence," but that such understandings develop over time—"out of sight, out of mind" may be a description of cognitive reality for infants. And, theories of memory including constructs such as short-term memory buffers are grounded in reliable data that people have major difficulty handling more than "the magic number seven plus or minus two" pieces of information in short-term memory. More broadly, general notions such as "schemata" are universal components of theories of memory. The initial findings, often obtained with very small *n*, have been replicated and extended numerous times.

It is, of course, impossible to prove that such claims actually hold for everyone. However, with precise enough definitions and operationalization of the research, replications of the studies can document the near-universality of the claims.

Beyond such cases, *caveat emptor* is probably the best attitude. A large number of claims appear to be universal, but they may need unpacking in various ways. Consider, for example, a generic claim for the effectiveness of instructional software: "One-on-one human tutoring is two standard deviations more effective than whole-class instruction. Our computer-based tutors are not yet that effective, but they are one standard deviation more effective than whole-class instruction." One can (and should) ask: effective according to what criteria? With what populations? Compared to what, under what circumstances? Absent compelling answers to such questions, there is reason to doubt the generality of the claims. Similarly, linguistic inflation and/or the desire for scientific prestige result in various claims regarding researchers having various theories or models. As discussed in Section IV, various theories (functionalism, conflict theory, symbolic interactionism, ethnomethodology, critical theory, ethnoscience, cognitive anthropology, exchange theory, psychodynamic theory, behaviorism, APOS theory, …) all have their applicability conditions. It is the responsibility of theorists to specify those conditions, to define the relevant constructs, and to address the limits (as well as the strengths) of what the theories can actually explain.

The same is true for the use of the term *mode*l. A model is more than a picture with a collection of objects and arrows. Claiming to have a model of a particular phenomenon means that one has specified particular objects and the relationships among them in the model, and that these entities correspond in some well-defined way to the objects and relationships in the phenomenon being modeled. That is a high standard. A cursory glance at any handbook related to education (e.g., Berliner & Calfee, 1996; Grouws, 1992; Kelley & Lesh, 2000; Sikula, 1996) reveals models galore. Let us examine a random example from each of these handbooks.

In the Berliner and Calfee *Handbook*, Mayer and Wittrock (1996) offer a model of the human information processing system in a schematic diagram (figure 3-1, p. 54) that includes inputs, outputs, and various kinds of memory. Various arrows go from one box to another. A key question (which the literature may well address, but which has to be asked of any such figure): Just what goes along the arrows? What *are* these processes called selecting, organizing, integrating, and storing, and how do they work?

In the Grouws *Handbook*, Romberg's (1992) chapter, "Perspectives on Scholarship and Research Methods," provides a "model for research and curriculum development" (figure 3-3, p. 52). Here too there are boxes and arrows, with arrows coming from the boxes labeled "classroom instruction" and "students' behaviors" to the box labeled "students' cognitions." Once again, the same question needs to be asked: just what goes along the arrows? And, what do the boxes really represent?

In the Kelley and Lesh (2000) *Handbook,* Lesh and Kelley (table 9.1, p. 198) describe a project in which (a) the goals for students include "constructing and refining models," (b) the teachers "construct and refine models to make sense of students' modeling activities," and (c) the "researchers develop models to make sense of the teachers' and students' modeling activities." Now, just what are the models in this case? What are the objects and relationships among them, and how do they correspond to the objects being modeled?

In the Sikula (1996) *Handbook*, Christensen reproduces two "teacher education design models" (figures 3.1 and 3.2) used by institutions of higher education to describe their teacher education programs. These almost defy description. The first is a Venn diagram (no arrows) in which the outer ring appears to be a "diverse global society," the next ring inward is labeled "private university/School of Education/Christian Environment," and the next ring contains "facilitator/lifelong scholar/professional/decisionmaker," inside of which are four interlocking rings. The second model appears in the outline of a tree, with "applied research," "professional societies," "world of practice," and "state guidelines" at its roots, and a series of arrows that ultimately arrive (via "program goals and objectives," general education," and more) to the "practicing professional." It seems, alas, that the seductions of scientism that led to the adoption of experimental paradigms in the 20th century lives on in the field's wish to claim "theories" and "models" as part of its working apparatus. The aspiration is admirable if and only if it is matched with a concomitant commitment to rigor.

VII. NOTES ON THE PREPARATION OF RESEARCHERS

Section III of this chapter discussed a series of assertions regarding desiderata for high quality research, among them the following:

- One must guard against the dangers of compartmentalization. Educators need a sense of the "big picture" and of how things fit together.
- One must guard against the dangers of being superficial. Generally speaking, high quality research comes when one has a deep and focused understanding of the area being examined.
- Researchers should be self-consciously aware of their theoretical perspectives and the entailments thereof. The methods they choose to employ should be selected on the basis of their appropriateness to address the questions that are considered important.
- Researchers must develop a deep understanding of what it means to make and justify claims about educational phenomena. What is a defensible claim? What is the scope of that claim? What kinds of evidence can be taken as legitimate warrants for that claim?

Much of the substance of this chapter has been devoted to addressing the substance of these last two points. The issues are by no means straightforward, even for established professionals. The question, then, is what can beginning researchers do in order to bootstrap some of the relevant knowledge?[15] I continue with additional assertions and some justifications for them.

- Students should have the opportunity to engage in research as early as possible in their careers, and they should be continually involved in various aspects of research—problem definition, methods selection, data gathering, and data analysis. Students should be encouraged, early on, to formulate problems and try to solve them (even if their first attempts are as awkward as a baby's first steps).[16]

The reason is simple: research is not a spectator sport and people will not develop a feel for doing research until they start doing it. This is the case even when one is learning to master standard techniques. It is especially the case when the research calls for the kinds of problem framing and methods development that are now part and parcel of our ongoing work. One colleague has summarized the issue succinctly as follows: "The best way to succeed is to fail early and often—with the appropriate support and guidance, of course." This chapter has emphasized the fact that there are myriad places where one can go wrong when doing research. Fundamental errors can occur in the ways one conceptualizes a problem, selects data, or analyzes them (to name just a few). Everyone will make mistakes. With the proper feedback and reflectiveness, one will learn from those mistakes. It makes sense, I believe, to start this process as early as possible. (To put things in very direct terms: Would we rather have a student make a major conceptual error in a course project or in pilot work for a thesis?)

- Multiple perspectives and multiple sources of feedback are good things. Students are likely to learn more if their work is commented on by more than one faculty member—especially if the faculty's expertise overlap and complement each other's.

This is, I hope, self-evident.

- Living in a research culture makes a difference—that is where habits of mind get shaped. Living in a research culture helps develop the kinds of breadth, depth, and multiple perspectives that are essential for the conduct of good research. It also provides important opportunities for the refinement of one's work.

There may well be "independent scholars" (in the sense of those whose ideas have sprung almost completely from within), but I suspect they are relatively few in number and that most scholars profit from sustained membership in a congenial intellectual community. My experience has been that there is no better way to have one's ideas shaped than to be a member of a community in which your ideas and ideas related to them are discussed. Sometimes the shaping is obvious: one walks out of a discussion with new or different thoughts as a result of the exchange. Sometimes the shaping is extremely subtle: I have realized, after the fact, that some of my ideas were, in important ways, the product of my environment. That is, I was most unlikely to have come up with some of the ideas I've come up with had I not been engaged in long-term conversations with particular colleagues, and influenced by their thinking.

Moreover, students can pick up many skills through discussions of others' work, before they are ready to grapple with big problems on their own. (See the discussion of research groups, below, for more detail.)

An active research culture also serves as a crucible for the refinement of work in progress. This can be the case for student papers, or student presentations at meetings—but it is also the case for my own work, this chapter being a case in point. I bring drafts of all of my papers to my research group, which does me the favor of questioning the work in careful detail. I profit every bit as much from these exchanges as my students do when their work is being discussed.

- Passion helps (when appropriately harnessed, of course). People do their best work when they care about what they do. A program that allows students to pursue their interests is likely to result in a higher degree of commitment and higher quality work than a program that does not.

This, too, should be self-evident.

If one accepts these assertions, there are various pragmatic ways to insure that students have such experiences early in their careers as developing researchers. Some such mechanisms are described in the balance of this section.

Project-based courses

Roughly half of the courses in our program (Education in Mathematics, Science, and Technology Program at the University of California, Berkeley) require students to conduct an empirical project of some significant scope. In such courses there is usually a heavy reading load "up front." In the middle of the term, the reading load lightens as students design and implement projects that are related to the course content. At the end of the course, the projects are used as vehicles to reflect on that content. (In addition, students can negotiate projects that meet the requirements for more than one course. This allows them to work on projects of substantial scope, and to get feedback from more than one faculty member.)

Course projects come in all shapes and sizes. The default option, which is actually exercised by very few students, is to replicate a study discussed in the course. Another option is to make a minor modification or extension of a study examined in the course. A third, and the one most frequently taken, is to find a phenomenon of interest and try to make sense of it. The data examined might be videotapes of a classroom or school, students' work on particular instructional materials or a computer program, people's "out loud" thoughts as they try to solve problems, or just about anything else. Almost anything related to the general content of a course is considered fair game.

Here are examples of student projects in a recent first year course.

One student had been working for some time as part of a team developing a test of "mathematical ability" that was being administered to thousands of students and analyzed using statistical measures. For her course project, she selected some people at various points on the spectrum from “ordinary beginner” to “talented expert”—the latter being a faculty member in mathematics. She hypothesized the kinds of performance that people with different levels of mathematical ability would display when they worked the problems, and then videotaped the people solving the problems. The reality of people’s performance was an eye-opener: some novices displayed much more effective problem solving practices on some of the problems than she expected, and her “expert” engaged in rather sloppy reasoning in places. This experience led her to question some of the assumptions she had been making about the problems, and about what people’s test scores really meant.

A second student hypothesized that girls and boys would act differently in same-sex problem solving groups than they would if all the other students in the group (of four or five) were of the opposite sex. As it happened, the people she chose (somewhat randomly) as the main subjects in her study tended to have robust character traits (shyness in some cases, aggressiveness in others), and there wasn’t much apparent difference in their performance. In reviewing the tapes, however, the student became interested in how collaborative the various groups were. She began to develop a coding scheme that looked at comments from students that invited reaction, versus those that were neutral or closed off conversation. This was a legitimate first step toward the quantification of collaborativeness—and a good example of problem definition and method creation (with help, of course).

A third student analyzed videotapes from an experimental course on mathematical representations that had been taught as part of a colleague’s R&D work the previous summer. As part of the project, the student in my course examined the beliefs of a student in the experimental course regarding what “counts” as being mathematical; he then tried to correlate the second student’s beliefs with her behavior. The student in the experimental course tended to disparage successful qualitative reasoning as “mere” common sense while giving high praise to mathematical behavior that included writing and solving equations—even though the equations she praised were (from our perspective) pretty much gobbledygook. Such beliefs seemed to play out in her actions during the course as well. (The evidence that my student offered in support of this claim was rather tenuous. That fact catalyzed some productive discussions about what it takes to justify such claims.)

Other projects dealt with student and teacher perceptions of a “reform” mathematics course, an attempt to analyze the teaching of a master teacher, the use of artifacts such as white boards (instead of individual sheets of paper) to catalyze interactions during group problem solving, and more.

How good were the projects? The truth is that when beginning students try to carry out such projects, their attempts tend to be seriously flawed. Students come to realize that they didn’t see the things they expected to see, that they can’t make the arguments they thought they’d be able to make … and sometimes, that there are interesting and unexpected leads in what they did see, which provide pointers to issues they’d like to pursue. Almost all of the papers were problematic in some way or other. That is no surprise; the students didn’t yet have the background to design or carry out near-perfect studies. Indeed, I think what happened is quite healthy. What the course offered was an institutionally supported way to make mistakes in the process of trying to define and work on a non-trivial research problem. Of course, this is only profitable if the students have the opportunity to learn from their mistakes. As part of their projects, students are asked to say how they would do things differently if they had them to do over again. (And—see the discussion of first and second year projects—they often have the opportunity to do them over again.) Then, in class discussion of the projects (which are presented formally as though at professional meetings) and in faculty evaluations of them, there is extended discussion of what worked, what didn’t, and what might be done about it.

Typically, students will take a number of courses each year that have such projects. In that way, the program offers an institutionalized mechanism for failing early and often—and for learning from those failures. They have the opportunity to develop their own perspectives on issues, refine their ideas and their methods, and try them out on critical but sympathetic audiences.

First and second year projects

The scope and quality of course projects are usually limited by the obvious constraint: things have to be done in the midst of one semester. For this reason course projects often have the character of pilot studies: an idea has been explored but there was not time to work it out right. To provide such opportunities we also require much more substantial project work.

In the summer following the first year of the program, and again in the summer following the second, students are required to conduct and write up more extensive studies. Typically, these first and second year projects are extensions of course projects: a course project may have yielded some tantalizing results, so the student goes back to gather more (or better) data to explore the issue in greater depth. With some frequency, project work is cumulative: a second year project is an outgrowth or modification of a first year project, and may itself evolve into a dissertation project. These projects are expected to meet rather stringent standards. They are to be written up as though for publication, and are judged accordingly. Each project report is read by two faculty members, and the discussion of the student's project is a major component of our annual student evaluation.

Even though they come on the heels of course work, first year projects can turn out to be seriously flawed, in which case the students are told to revise them and try again. Many are respectable, however, and only need minor revisions. Either way, it is healthy to establish a high standard for judgment and to provide rigorous feedback. Second year projects tend to be of uniformly high quality, and a fair number of them have been published; a significant number are presented at professional meetings. The acceptance rate for student proposals and papers is quite high. I have no doubt that the students' success is attributable to the fact that we provide them with consistent opportunities do independent work and to receive critical feedback on it.

Research groups

As noted above, I believe that students are more likely to become productive researchers, and to develop useful habits and perspectives more rapidly, if they are members of a research *community*. When you are constantly engaging with people who live and breathe research issues, participating in the development of their ideas and in their successes and failures, you are much more likely to pick up "what counts" than you would be if you were working in isolation.

In our program, each faculty member has at least one research group. Every student in the program is expected to participate regularly in one or more research groups. Many students attend two or more groups, because they find the complementary perspectives and expertise to be valuable.

While there is tremendous variation, certain properties tend to be present if a research group or community is functioning well. Three of those properties are as follows:

- There is a sense of purpose and meaningfulness: much of the work done really matters to the people involved. (Work is not seen as busy work, but as part of what needs to get done to advance the enterprise.)
- Much of the work being done is *visible*—the processes of doing research, including mulling through problems, are public property in the sense that dilemmas are shared and community input is valued as a way of solving them. There is a culture of reflectiveness, where the expectation is that problematic issues will be raised, and that members of the community will consider contributing to their solution (even problems whose solution does not contribute to their own progress) as one of their communal responsibilities. The culture is such that there is room for the work and contributions of all members to be taken seriously.

As I mentioned above, I bring my own writings to the group for critiques. My students have commented that this has played a significant role in demystifying the writing process. When they see me struggling to express myself, and they see the number of drafts I work through (they commented on draft K of this chapter—I work my way through the alphabet!), they get a better sense of what it takes to produce a polished paper. Likewise, students have seen me struggle with new ideas for my ongoing work—and have played a significant role in shaping it.

- The work and interactions of the group provide a series of "handholds" that allow individuals at various levels of knowledge and expertise to contribute meaningfully to the enterprise, and to make parts of it their own. Newcomers' contributions may consist of routine work in the service of the cause (e.g., first-year students in a research group that has an ongoing development project might play a small role in the development process, help field test some materials, or help videotape lessons). At the same time, those students are present for the theoretical discussions and are invited to contribute whenever they felt comfortable doing so. Typically, early contributions consist of occasional comments or questions, as beginners try to sort out the spirit or the details of what is being done. As they become more central members of the community, the character of their questions and contributions tends to evolve. The students are likely to take on larger tasks, individually or in collaboration, and they increasingly take on ownership of tasks and ideas.

[Another way to describe this process in somewhat more theoretical terms is that a functioning research community provides multiple opportunities for legitimate peripheral participation. As once-peripheral members become more central to the enterprise they find more means of achieving centrality, and there is room and access for new members at the periphery. The detailed examination of this process would be a most welcome study.]

There is, it should be stressed, no one model of a productive research group or community. Such communities may be very small, consisting of one senior researcher and a few students, or they may be rather large, including a substantial number of people with varied levels of skill and expertise. Moreover, no such group is static: depending on people involved and the tasks at hand (is a major focus of the group conceptualizing a new project, building a collaboration, "engineering change," designing or implementing materials, gathering data, analyzing data, writing or revising papers or proposals, or … ?), the day-to-day transactions of the group and its level of activity will vary. Among the activities that research groups in our environment have supported are the following:

Participation (whether as central player or legitimate peripheral participant) in a major ongoing project

The benefits of this kind of engagement were discussed immediately above.

Providing group members feedback on issues of importance to them.

One function of a research group is to serve as a critically supportive environment for discussions of student work. What is brought to the group can vary substantially. A student may have a vague idea for a project and ask for the group's help in honing that idea. He or she may have some data and want to see the group's reactions, or may want to see the group's reaction to a tentative explanation of those data. The student may have a draft piece of work—a course project, a master's thesis, a dissertation proposal, a chapter of a dissertation, a proposal for a conference presentation, or a paper for submission—and want feedback. Sessions are scheduled with enough lead time so that group members are expected to go through the relevant materials, and to serve as colleagues in providing help to the presenter.

I note that it is not at all necessary for the students to be working on the "same thing" in order for them to take each other's work seriously. For example, students in one research group were working simultaneously on transfer, teacher knowledge, cultural forces shaping the effectiveness of instruction, and issues of reflection on professional growth and integration. Yet discussions of these students' ongoing work—from the early stages of problem formulation through the stage of selecting data, agonizing over what the data meant, and then writing things up in ways that were cogent and compelling, all proceeded in parallel and in comfort. What made the group function effectively was a common interest in helping each other work things through, and an understanding that at some fundamental level everyone was grappling with the same issues. No matter whose work was being discussed, conversations were all grounded in the same kinds of questions: What are you trying to say (what are the "punch lines")? Why would anyone think this is important? What kind of evidence will convince people that what you are saying is justified? What are counter-interpretations? What position will you be in if the data don't tell the story you'd like? What are the implications of your expected results, and why should anyone believe them? Of course the group tried not only to raise the questions, but also to help answer them.

It should be noted that in these conversations, everyone profits. The presenter gets feedback that is useful. The others hone their skills in understanding and critiquing research, and in learning to ask others the kinds of questions they will have to ask themselves as independent researchers.

Dealing with topics or readings of interest

Research groups often serve as reading or discussion groups. This provides a way to delve deeply into issues as a community. Groups have, at various times, decided to "go to school" on various theoretical perspectives (constructivism, situated cognition), to explore the strengths and limitations of particular research methods, or to discuss papers on topics that just plain seemed interesting.

Providing a critical but friendly audience for practice talk

Prior to major professional meetings, research groups often provide forums for practice presentations. In most groups, students and faculty rehearse their presentations before the group before they go "public." It is much better to learn to deal with tough questions in the comfort of a research group than to hear them for the first time when at the podium in a public presentation!

What really matters in all of the above? What counts from my perspective is to provide a supportive environment that lives and breathes research issues, that is open and reflective, that allows people to pursue ideas that they really care about, and that provides them with many opportunities to learn, early on, from the mistakes they will inevitably make.

In closing this section, I would like to address an issue that Frank Lester raised when he reviewed a draft of this chapter:

> I would like to read about what an overall program might look like at three types of institutions: (1) those few that expect students to begin thinking seriously about research from the beginning, (2) those that are preparing math educators who might also do some research, but who surely will be (primarily) consumers of research, and (3) those that simply require students to write a dissertation as a final requirement for the terminal degree. My fear is that institutions in the third category are preparing most of our future math educators. Even if this is not the case, it surely is true that there are relatively few category 1 institutions. In fact, I would like to see him discuss the type of program appropriate for category 2 institutions and to engage in some speculation about how to prepare math educators to be good consumers and interpreters of research.

I have, elsewhere (Schoenfeld, 1999b) discussed ways to think about core content for a doctoral program in mathematics education; what follows are "headlines" of that discussion. First, content. There is no solution to the "content problem," but one can satisfice—the goal is to give students a sense of the many flavors of educational work and their contributions. There are, I think, reasonable approaches grounded in the structure of any institution. On the one hand, one needs to provide "disciplinary" information. This can be done via core courses that reflect the disciplinary organization of the institution. For example, Berkeley's School of Education is organized into three overarching academic units called "Areas." Faculty in each Area are encouraged to propose core courses that give a taste of mainstream issues, perspectives, and methods in that area, while highlighting connections with and perspectives and methods from the other areas. (If a school has *n* areas where *n* is large, courses bridging such areas can be co-taught.) On the other hand, students should come to understand how educational issues transcend disciplinary boundaries. One way to do this is to offer a series of courses on "cross-cutting topics," where faculty from different units bring varied disciplinary perspectives to the study of such issues. Any of a number of topics—teacher preparation, assessment, or diversity, to name three—can serve as topics for discussion.

Second, methods. It is impossible in a few introductory courses to provide the depth and breadth of coverage that will result in students being adequately prepared for the research they will do. We can prepare students to be knowledgeable and skeptical consumers, and we can help engender in them an understanding of the fundamental issues. We can, in the best of all possible worlds, help them develop the right kinds of questioning attitudes—asking, as they proceed with their work, what they can say with justification, and how best to approach issues so they can make strongly warranted claims. But we should not make the mistake of thinking that methods courses will prepare students adequately for their research. If students emerge from their methods courses with a sense of how to approach a problem, of how to select methods that seem reasonable, and of where to go for help when they realize the limits of those methods, then the core has been quite successful. My bias is that "less is more:" a small number of cases carefully studied will be more productive in the long run than an encyclopedic treatment (with each topic studied in the depth of a typical encyclopedia article). *Really* learning about methods should come when students try to use them—in courses, in projects, and as members of a research community.

Above and beyond core content, I would strongly recommend a program that contains a large number of project-based courses, and that has projects similar in kind to the first and second year projects discussed above. Research groups are extremely valuable, but if the environment does not support them, some of their functions can be achieved through research seminars or through modifications of the core courses.

This discussion may not seem fully responsive to the issues raised above by Frank Lester, in that I do not separate out three different kinds of institutions as he suggests. My failure to do so is deliberate. Frank is right that at present there are relatively few institutions at which students begin conducting research early in their careers. From my perspective, that is most unfortunate—even if most students at a particular institution intend to become consumers rather than producers of research. One learns a great deal about how research is done—about what to believe, what not to believe—by trying to conduct it and learning from the experience.

A case in point is the first student whose work was discussed in the section on "project-based courses." Her course project consisted of taking a close look at what people of (ostensibly) different mathematical abilities actually did when they worked mathematics problems. I indicated above that by virtue of having conducted the project, the student developed a much more nuanced view of problem solving abilities, and of what tests reveal about such competencies.

As it happens, that student did not intend to have a career as a researcher. She was enrolled in our teacher preparation program and she went on, as planned, to become a teacher. Her presence in the project-based course was no accident. Our teacher preparation program is designed so that student teachers and beginning doctoral students are enrolled in many of the same courses. Student teachers are also enrolled in faculty's research groups. The idea is that this kind of hands on experience with research will enable them to develop a much better understanding of thinking, teaching, and learning than they would by merely reading about it. Even if they never go on to do formal research on their own, they will be much better consumers of research, and they will be better at understanding student thinking for having explored it in detail. I think *all* mathematics educators should have such experiences, no matter what their intended careers.

VIII. A FINAL COMMENT

As one reflects on the state of the field, it is worth recalling the historical data with which this chapter began. Mathematics education began to coalesce as a discipline only a few decades ago, with its first professional meetings and journals appearing in the late 1960s and early 1970s. Its growth since then has been nothing short of phenomenal. Once held tightly in the stranglehold of a reductive epistemology and scientistic methods, the field has blossomed. Important phenomena as diverse as "metacognition" and "communities of practice" have been uncovered and elaborated in substantial detail. Important theoretical frameworks as diverse as cognitivism, ethnomethodology, and critical theory (to name just a few) have been developed. Methods as diverse as cognitive modeling and discourse analysis have been crafted, and approaches for the principled design of educational materials, such as design experiments, have been developed. The results have been a broad range of thought-provoking interpretations of mathematical thinking and learning. During one scholar's lifetime, the field has progressed from the point where controlled laboratory studies were necessary to explore simple cognitive phenomena to the point where the detailed modeling of thinking and learning in complex social environments is possible.

At the same time, the field confronts at least two major difficulties. First, much of the growth has been chaotic. As is absolutely characteristic of young fields experiencing rapid growth, much of the early work has been revealed to be seriously flawed. As discussed above, unarticulated theoretical biases or unrecognized methodological difficulties undermined the trustworthiness of a good deal of work that seemed perfectly reasonable at the time it was done. This should not cause hand wringing—such is the nature of the enterprise—but it should serve as a stimulus for devoting seriously increased attention to issues of theory and method. As the field matures, it should develop and impose the highest standards for its own conduct.

Second, in the United States at least, there is the serious risk of a return to scientism in the name of "science" and under the banner of the "gold standard" of randomized controlled trials. I hope to have made clear in this chapter that the use of quantitative methods, while apparently more straightforward than the use of qualitative methods, is in fact every bit as complex—and every bit as liable to misinterpretation. The challenge, no matter what kind of method is being used, is to use it properly—to gather evidence that provides solid warrants for the claims being made. My hope is that the framework discussed in Section VI and the criteria discussed in Section V will prove useful tools along these lines.

NOTES

1. These are problems stated in forms such as

$$\begin{array}{cccccc} & & S & E & N & D \\ & + & M & O & R & E \\ \hline = & M & O & N & E & Y \end{array} \quad \text{or} \quad \begin{array}{ccccccc} & & D & O & N & A & L & D \\ & + & G & E & R & A & L & D \\ \hline & = & R & O & B & E & R & T. \end{array}$$

   A solution to a problem consists of replacing each letter in the given form with a unique digit from 0 to 9 so that when all the replacements are made, the arithmetic sum that results is correct.
2. Vygotsky died in 1934, so the roots of this work extend quite deeply. The 1962 and 1978 dates of publication of *Thought and Language* and *Mind in Society* represent the appearance of his work in English translation.

3. The discussion of issues 1 and 2 in this section is a brief reprise of an argument made in Schoenfeld (1999b); see that paper for more extensive detail. The discussion of issue 3 is taken, with slight modifications, from Schoenfeld (1999a).
4. A caveat: the claim is that a significant proportion of educational research *can* (and when possible, should) be carried out in "real" contexts. However, at different points in the development of a field, it may be difficult for any one corpus of work to contribute simultaneously to both theory and practice. Sometimes the state of theory is such that it may best be nurtured, temporarily, aside from significant considerations of use (consider the origins of cognitive science, which was nurtured in laboratory studies). Sometimes the need to solve practical problems seems so urgent that theoretical considerations may be given secondary status (consider the post-Sputnik period, during which engineering efforts such as "putting a man on the moon" took priority). Figure 19.3 should be taken as a heuristic guide, with the upper right-hand quadrant representing a desirable site for work, when possible.
5. What follows is complex, but perhaps not complex enough. In Figure 19.5 each of the boxes is static and each of the arrows is unidirectional. In reality, of course, the process of data interpretation is dynamic: conceptual models and representational systems evolve as one comes to a better understanding of the relevant phenomena, and the process is dialectic rather than linear. Readers who wish to wallow in the complexities of the research process, among other things, may wish to explore Latour (1988, 1999).
6. This discussion is expanded from Schoenfeld (2000).
7. 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.
8. "Folk wisdom" is a case in point. *Everything* can be explained (at least post hoc) by folk wisdom. Depending on circumstances, for example, you can invoke the maxim "haste makes waste" to say that things must be done slowly and carefully, or "a stitch in time saves nine" to say that being speedy is essential. A "theory" that explains everything explains nothing.
9. I understand that this appears to be a strongly Popper-like (1963) stance, and that alternative stances such as those taken by Toulmin (1958) or Pickering (1995) are more contextual in character. While acknowledging the necessity for context-based theories, I think that one can look for aspects of falsifiability (even in particular contexts or in rough equivalence classes of them) without being committed to a fully Popper-like stance. See the discussion of replicability in the following section.
10. It is possible to perform calculation such as 384 × 673 mentally by rehearsing the subtotals. For example, one can calculate $3 \times 384 = 1152$ and repeat "1152" mentally until it becomes a "chunk" which only occupies one short-term memory buffer. "Chunking" is a well-documented mechanism by which people can perform mental tasks.
11. The hypothetical limit of the number of short-term memory buffers would be of little interest if it applied only to tasks such as multiplication. Miller's finding, however, has great scope: there is a wide range of tasks, from very many domains, on which people begin to falter badly when the number of things they have to "keep in mind" approaches seven.
12. In simplest terms, one group of students used what is called the "count all" strategy for addition, while the second group had developed the "counting on" strategy.
13. These categories are not crisply defined, of course; the character of the event is a function of the perspective of the researcher. For example: from the perspective of teacher-researchers involved in implementing a new curriculum, their work is an attempt at change. From the perspective of anthropologists examining the "cultures" of their classrooms, the observations may be "descriptions of the reality of a school in flux." Both perspectives on the same set of events are possible.
14. This is not a hypothetical issue. In my book *Mathematical Problem Solving* (Schoenfeld, 1985), I describe a scheme for analyzing transcripts of problem solving sessions that focuses on "make or break" decisions during problem solving. Following the book's publication, I received a substantial number of communications from colleagues who said the scheme had not helped them analyze transcripts of students solving "problems" such as finding the product of two three-digit numbers. It should have been no surprise that strategic decisions are few and far between when one is working on problems that are purely procedural.
15. The discussion that follows is distilled from the concluding sections of Schoenfeld, 1999b.
16. The statement reveals a personal bias, that "problem-driven" research (rather than "method-driven" research and to some degree, "theory-driven research") is the most profitable way for the field to progress at present. When theory is stable and methods are well established, fields can progress by "working out the details"—using a standard set of methods to obtain results. When theories and methods are unstable, however, a profitable strategy is often to select problems that are of theoretical and pragmatic interest (recall Pasteur's quadrant), and that have the potential to be solvable.

Working out the solutions, often through adaptations of known methods, can contribute to the development of theory while expanding the community's methodological tool kit.

REFERENCES

American Association of University Women. (1992). *The AAUW report: How schools shortchange girls.* Annapolis Junction, MD: The AAUW Educational Foundation.

Artigue, M. (1999). The teaching and learning of mathematics at the university level: Crucial questions for contemporary research in education. *Notices of the American Mathematical Society,* (46), 1377–1385.

Asiala, M., Brown, A., de Vries, D., Dubinsky, E., Mathews, D., & Thomas, K. (1996). A framework for research and curriculum development in undergraduate mathematics education. In J. Kaput, A. Schoenfeld, & E. Dubinsky (Eds.), *Research in collegiate mathematics education, Vol. II* (pp. 1–32). Washington, DC: Conference Board of the Mathematical Sciences.

Ausubel, D. P. (1968). *Educational psychology: A cognitive view.* New York: Holt, Rinehart, & Winston.

Ball, D., & Lampert, M. (1999). Multiples of evidence, time, and perspective: Revising the study of teaching and learning. In: E. Lagemann & L. Shulman (Eds.), *Issues in education research: Problems and possibilities* (pp. 371–398). New York: Jossey-Bass.

Bauersfeld, H. (1980). Hidden dimensions in the so-called reality of a mathematics classroom. *Educational studies in mathematics, 11*(1), 109–136.

Bauersfeld, H. (1993). Theoretical perspectives on interaction in the mathematics classroom. In R. Bieler, R. Scholz, R. Strasser, & B. Winkelmann (Eds.), *Didactics of mathematics as a scientific discipline* (pp. 133–146). Dordrecht, Netherlands: Kluwer.

Bauersfeld, H. (1995). "Language games" in the mathematics classroom: their functions and their effects. In P. Cobb & H. Bauersfeld (Eds.), *The emergence of mathematical meaning* (pp. 271–292). Mahwah, NJ: Erlbaum.

Beeby, T., Burkhardt. H., & Fraser, R. (1979). *Systematic Classroom Analysis Notation (SCAN) for mathematics lessons.* Nottingham, UK: Shell Centre for Mathematical Education.

Berliner, D., & Calfee, R. (Eds.). (1996). *Handbook of educational psychology.* New York: MacMillan.

Boaler, Jo. (2002). *Experiencing school mathematics (Revised and expanded edition).* Mahwah, NJ: Erlbaum.

Brophy, J., & Good, T. (1986). Teacher behavior and student achievement. In M. C. Wittrock (Ed.), *Handbook of Research on Teaching* (3rd ed., pp. 328–375). New York: Macmillan.

Brousseau, G. (1997). *Theory of didactical situations in mathematics: Didactique des mathematiques, 1970–1990.* Dordrecht, Netherlands: Kluwer.

Brown, A. (1992) Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *Journal of the Learning Sciences, 2*(2), 141–178.

Brown, A., & Campione, J. (1996) Psychological theory and the design of innovative learning environments: On procedures, principles, and systems. In L. Schauble & R. Glaser (Eds.), *Innovations in learning: New environments for education* (pp. 289–325). Mahwah, NJ: Erlbaum.

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

Brueckner, L. J., & Kelley, F. (1930). A critical evaluation of methods of analyzing practice in fractions. In G. M. Whipple (Ed.), *Report of the Society's committee on arithmetic (The twenty-ninth Yearbook of the National Society for the Study of Education)* (pp. 525–534). Bloomington, IL: Public School Publishing Company.

Bush, V. (Reprinted 1990) *Science, the endless frontier. A report to the President on a program for postwar scientific research.* Washington, DC: National Science Foundation.

Buswell, G. (1951). The psychology of learning in relation to the teaching of arithmetic. In N. B. Henry (Ed.), *The teaching of arithmetic (The fiftieth Yearbook of the National Society for the Study of Education, Part II)* (pp. 143–154). Chicago: University of Chicago Press.

Campbell, D. T., & Stanley, J. C. (1966) *Experimental and quasi-experimental designs for research.* Boston: Houghton Mifflin.

Carraher, D. W. (1991). Mathematics in and out of school: A selective review of studies from Brazil. In M. Harris (Ed.), *Schools, mathematics, and work* (pp. 169–201). London: Falmer.

Christensen, D. (1996). The professional knowledge-research base for teacher education. In J. Sikula (Ed.), *Handbook of research in teacher education* (3rd ed.) (pp. 38–52). New York: Macmillan.

Clapp, F. (1926/1995). Some recent investigations in arithmetic. In *A general survey of progress in the last twenty-five years (First Yearbook of the National Council of Teachers of Mathematics)* (pp. 166–185). Reston, VA: NCTM.

Clement, J. (1982). Algebra word problem solutions: thought processes underlying a common misconception. *Journal for Research in Mathematics Education, 13*(1), 16–30.

Clement, J., Lochhead, J., & Monk, G. S. (1981). Translation difficulties in learning mathematics. *American Mathematical Monthly, 88*(3), 286–290.

Clement, J. (2000). Analysis of clinical interviews: Foundations and model viability. In A. E. Kelley & R.d A. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 547–590). Mahwah, NJ: Erlbaum.

Cobb, P. (1995). Mathematical learning and small-group interaction: four case studies. In P. Cobb & H. Bauersfeld (Eds.), *The emergence of mathematical meaning* (pp. 25–130). Mahwah, NJ: Erlbaum.

Cobb, P. (2000). Conducting teaching experiments in collaboration with teachers. In A.y E. Kelley & R. A. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 307–334). Mahwah, NJ: Erlbaum.

Cognition and technology group at Vanderbilt. (1997). *The Jasper project: Lessons in curriculum, instruction, assessment, and professional development.* Mahwah, NJ: Erlbaum.

Cohen, D. (1990). A revolution in one classroom: The case of Mrs. Oublier. *Education Evaluation and Policy Analysis, 12*(3), 311–329.

Cooney, T. (1985). A beginning teacher's view of problem solving. *Journal for Research in Mathematics Education, 16*, 324–336.

Eisenhart, M., Borko, H., Underhill, R., Brown, C., Jones, D., & Agard, P. (1993). Conceptual knowledge falls through the cracks: Complexities of learning to teach mathematics for understanding. *Journal for Research in Mathematics Education, 24,* 8–40.

English, L. D. (Ed.). (1997). *Mathematical reasoning: analogies, metaphors, and images.* Mahwah, NJ: Erlbaum.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87*, 215–251.

Evertson, C., Emmer, E., & Brophy, J. (1980). Predictors of effective teaching in junior high mathematics classrooms. *Journal for Research in Mathematics Education, 12*, 167–178.

Fawcett, H. (1938). *The nature of proof. (Thirteenth Yearbook of the National Council of Teachers of Mathematics.)* New York: Teachers College.

Freudenthal, H. (1973). *Mathematics as an educational task.* Dordrecht, Netherlands: Reidel.

Freudenthal, H. (1983). *Didactical phenomenology of mathematical structures.* Dordrecht, Netherlands: Reidel.

Gardner, H. (1987). *The mind's new science: A history of the cognitive revolution.* New York: Basic Books.

Gee, J., Michaels, S., & O'Connor, M. C. (1992). Discourse analysis. In M. LeCompte, W. Millroy, & J. Preissle (Eds.), *Handbook of qualitative research in education* (pp. 227–292). New York: Academic Press.

Glaser, B., & Strauss, A. (1967) *The discovery of grounded theory.* London: Weidenfeld & Nicholson.

Glaser, R., & Linn, R. (Eds.). (1997). *Assessment in transition: Monitoring the nation's educational progress.* Stanford, CA: National Academy of Education.

Green, J. L., Camilli, G., & Elmore, P. B. (Eds.) (2006). *Handbook for complementary methods in education research.* Mahwah, NJ: Erlbaum.

Greeno, J. G., & The Middle-school Mathematics through Applications Project Group. (1997). Theories and practices of thinking and learning to think. *American Journal of Education, 106*, 85–126.

Greeno, J. G., & the Middle-School Mathematics through Applications Project Group. (1998). The situativity of cognition, learning, and research. *American Psychologist, 53*, 5–26.

Greeno, J. G., Pearson, P. D., & Schoenfeld, A. H. (1997). Implications for the National Assessment of Educational Progress of Research on Learning and Cognition. In R. Glaser & R. Linn (Eds.), *Assessment in transition: Monitoring the nation's educational progress, background studies* (pp. 152–215). Stanford, CA: National Academy of Education.

Grouws, Douglas A. (Ed.). (1992). *Handbook of research on mathematics teaching and learning.* New York: Macmillan.

Hadamard, J. (1945). *An essay on the psychology of invention in the mathematical field, by Jacques Hadamard.* Princeton, NJ: Princeton University Press.

Hardy, G. H. (1967). *A mathematician's apology.* Cambridge: Cambridge University Press.

Heath, S. B. (1999). Discipline and disciplines in educational research: Elusive goals? In E. Lagemann & L. Shulman (Eds.), *Issues in education research: Problems and possibilities* (pp. 203–223). New York: Jossey-Bass.

Kelley, A. E. (2003, January/February). (Guest Ed.) Theme issue: The role of design in educational research. *Educational Researcher* 32(1).

Kelley, A. E., & Lesh, R. A. (2000). *Handbook of research design in mathematics and science education.* Mahwah, NJ: Erlbaum.

Kilpatrick, J. (1978). Variables and methodologies in research on problem solving. In L. L. Hatfield (Ed.), *Mathematical problem solving* (pp. 7–20). Columbus, OH: ERIC.

Kilpatrick, J., & Wirszup, I. (Eds.). (1975). *Soviet studies in school mathematics.* Chicago: University of Chicago Press.

Knight, F. B. (1930). Introduction. In Guy M Whipple (Ed.), *Report of the Society's committee on arithmetic* (*The twenty-ninth yearbook of the National Society for the Study of Education*) (pp. 1–8). Bloomington, IL: Public School Publishing Company.

Krutetskii, V. I. (1976). *The psychology of mathematical abilities in school children.* J. Kilpatrick & I.Wirszup (Eds.), J. Teller (Trans.) Chicago: University of Chicago Press.

Kuhn, T. S. (1970). *The structure of scientific revolutions* (*2nd ed.*). Chicago: University of Chicago press.

Latour, Bruno. (1988) *Science in action: How to follow scientists and engineers through society.* Cambridge, MA: Harvard University Press.

Latour, Bruno. (1999) *Pandora's hope: Essays on the reality of science studies.* Cambridge, MA: Harvard University Press.

LeCompte, M., Millroy, W., & Preissle, J. (Eds.). (1992). *Handbook of qualitative research in education.* New York: Academic Press.

LeCompte M. & Preissle, J. (1993). *Ethnography and qualitative design in educational research* (2nd ed.). San Diego: Academic Press.

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

Lesh, R., & Kelley, A. E. (2000). Multitiered teaching experiments. In A. E. Kelley & R. Lesh, *Handbook of research design in mathematics and science education* (pp. 197–230). Mahwah, NJ: Erlbaum.

Lincoln, Y., & Guba, E. (1985). *Naturalistic inquiry.* Beverly Hills, CA: Sage.

Lucas, J., Branca, N., Goldberg, D., Kantowsky, M. G., Kellogg, H. & Smith, J. P. (1980). A process-sequence coding system for behavioral analysis of mathematical problem solving. In G. Goldin & E. McClintock (Eds.), *Task variables in mathematical problem solving* (pp. 353–372). Columbus, OH: ERIC.

Ma, L. (1999). *Knowing and teaching elementary mathematics: Teachers' understanding of fundamental mathematics in China and the United States.* Mahwah, NJ: Erlbaum.

Mayer, R., & Wittrock, M. (1996). Problem-solving transfer. In D. Berliner & R. Calfee (Eds.), *Handbook of educational psychology* (pp. 47–62). New York: MacMillan.

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

Moschkovich, J., & Brenner, M. E. (2000). Integrating a naturalistic paradigm into research on mathematics and science cognition and learning. In A. E. Kelley & R. A. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 457–486). Mahwah, NJ: Erlbaum.

Nasir, N., Cobb, P. (Eds.). (2002). Special Issue: Diversity, equity, and mathematical learning. *Mathematical Thinking and Learning, 4*(2&3).

National Council of Teachers of Mathematics (1926/1995). *A general survey of progress in the last twenty-five years (First Yearbook).* Reston, VA: NCTM.

Newell, A., & Simon, H. A. (1972). *Human problem solving.* Englewood Cliffs, NJ: Prentice Hall.

Nisbett, R. E., & T. Wilson. (1977). Telling more than we know: verbal reports on processes. *Psychological Review, 84*, 231–260.

Norem, G. M., & Knight, F. B. (1930). The learning of the one hundred multiplication combinations. In G. M Whipple (Ed.), *Report of the Society's committee on arithmetic* (*The twenty-ninth yearbook of the National Society for the Study of Education*) (pp. 551–568). Bloomington, IL: Public School Publishing Company.

Peters, R. S. (1965). *Brett's history of psychology* (*Revised and abridged by R. S. Peters*) (2nd ed.). Cambridge, MA: MIT Press.

Piaget, J. (1956). *The child's conception of space.* London: Routledge & Kegan Paul.

Piaget, J. (1969a). *The child's conception of number.* London: Routledge & Kegan Paul.

Piaget, J. (1969b). *The child's conception of time.* London: Routledge & Kegan Paul.

Piaget, J. (1970). *Genetic epistemology* (E. Duckworth, Trans.). New York: W.W. Norton.

Pickering, A. (1995). *The mangle of practice: Time, agency, and science.* Chicago: University of Chicago Press.

Poincare, H. (1913). *The foundations of science* (G. H. Halstead, Trans.). New York: Science Press.

Popper, K. (1963). *Conjectures and refutations: The growth of scientific knowledge.* London: Routledge & Kegan Paul.

Repp, A. C. (1930). Mixed versus isolated drill organization. In G. M Whipple (Ed.), *Report of the Society's committee on arithmetic* (*The twenty-ninth yearbook of the National Society for the Study of Education*) (pp. 535–550). Bloomington, IL: Public School Publishing Company.

Resnick, L. B. (1983). Toward a cognitive theory of instruction. In S. Paris, G. M. Olson, & H. W. Stevenson (Eds.), *Leaning and motivation in the classroom* (pp. 5–38). Hillsdale, NJ: Erlbaum.

Romberg, T. (1992). Perspectives on scholarship and research methods. In D. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 49–64). New York: Macmillan.

Rosnick, P., & Clement, C. (1980). Learning without understanding: the effect of tutoring strategies on algebra misconceptions. *Journal of Mathematical Behavior, 3*(1), 3–27.

Russell, B., & Whitehead, A. N. (1960). *Principia mathematica* (2nd ed.). Cambridge: Cambridge University Press.

Schauble, L., & Glaser, R. (1996). *Innovations in learning: New environments for education.* Mahwah, NJ: Erlbaum.

Schoenfeld, A. H. (1985). *Mathematical problem solving.* Orlando, FL: Academic Press.

Schoenfeld, A. H. (1989). Explorations of students' mathematical beliefs and behavior. *Journal for Research in Mathematics Education, 20*(4), 338–355.

Schoenfeld, A. H. (Ed.) (1992). *Research methods in and for the learning sciences.* A special issue of *The Journal of the Learning Sciences, 2*, 2.

Schoenfeld, A. H. (1994). A discourse on methods. *Journal for Research in Mathematics Education, 25*(6), 697–710.

Schoenfeld, A. H. (1998). Toward a theory of teaching-in-context. *Issues in Education*, *4*(1), 1–94.

Schoenfeld, A. H. (1999a). Looking toward the 21st century: Challenges of educational theory and practice. *Educational researcher*, *28*(7), 4–14.

Schoenfeld, A. H. (1999b). The core, the canon, and the development of research skills: Issues in the preparation of education researchers. In E. Lagemann & L. Shulman (Eds.), *Issues in education research: Problems and possibilities* (pp. 166–202). New York: Jossey-Bass.

Schoenfeld, A. H. (1999c) (Special Issue Editor). *Examining the complexity of teaching.* Special issue of the *Journal of Mathematical Behavior*, *18*(3).

Schoenfeld, A. H. (2000) Purposes and methods of research in mathematics education. *Notices of the American Mathematical Society, 47*(6), 2–10.

Schoenfeld, A. H. (2002a) A highly interactive discourse structure. In J. Brophy (Ed.), *Social sonstructivist teaching: Its affordances and constraints* (Vol. 9 of the series *Advances in Research on Teaching*) (pp. 131–170). New York: Elsevier.

Schoenfeld, A. H. (2002b). Research methods in (mathematics) education. In L. English (Ed.), *Handbook of international research in mathematics education* (pp. 435–488). Mahwah, NJ: Erlbaum.

Schoenfeld, A. H. (2006) Design experiments. In J. L. Green, G. Camilli, & P. B. Elmore (Eds.) *Handbook for complementary methods in education Research* (pp. 193–206). Mahwah, NJ: Erlbaum.

Schoenfeld, A. H. (2006, March). What doesn't work: The challenge and failure of the What Works Clearinghouse to conduct meaningful reviews of studies of mathematics curricula. *Educational Researcher*, *35*(3), 13–21.

Schoenfeld, A. H. (Ed.) (in press). A study of teaching: Multiple lenses, multiple views. *Journal for Research in Mathematics Education* Monograph Series.

Schorling, R. (1926/1995). Suggestions for the solution of an important problem that has arisen during the last quarter of a century. In *A general survey of progress in the last twenty-five years* (First *Yearbook* of the National Council of Teachers of Mathematics) (pp. 166–185). Reston, VA: NCTM.

Sfard, A. (1994). Reification as the birth of metaphor. *For the Learning of Mathematics, 14*(1), 44–54.

Sfard, A., & McClain, K. (2002). Special Issue: Analyzing tools: Perspectives on the role of designed artifacts in mathematics learning. *Journal of the Learning Sciences, 11*(2 & 3).

Sherin, M., & Sherin, B. (in press). Moving from shared data to shared frameworks. In A. H. Schoenfeld (Ed.), A Study of Teaching: Multiple Lenses, Multiple Views. *Journal for Research in Mathematics Education* Monograph Series.

Shirk, G. B. (1972). *An examination of the conceptual frameworks of beginning mathematical teachers.* Unpublished dissertation. Urbana-Champaign: University of Illinois.

Sikula, J. (Ed.). (1996). *Handbook of research on teacher education* (2nd ed.). New York: Macmillan.

Skinner, B. F. (1958). Teaching machines. *Science, 128*, 969–977.

Steffe, L., Nesher, P., Cobb, P., Goldin, G., & Greer, B. (1996). *Theories of mathematical learning.* Mahwah, NJ: Erlbaum.

Steffe, L., & Thompson, P. (2000). Teaching experiment methodology: Underlying principles and essential elements. In A. E. Kelley & R. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 267–306). Mahwah, NJ: Erlbaum.

Stigler, J, & Hiebert, J. (1999). *The teaching gap.* New York: Free Press.

Stokes, D. E. (1997). *Pasteur's quadrant: Basic science and technical innovation.* Washington, DC: Brookings.

Thorndike, E. L. (1922). *The psychology of arithmetic.* New York: Macmillan

Toulmin, S. (1958). *The uses of argument.* Cambridge: Cambridge University Press.

Vygotsky, L. S. (1962). *Thought and language.* Cambridge, MA: MIT Press.

Vygotsky, L. S. (1978). *Mind in society: the development of higher psychological processes.* Cambridge, MA: Harvard University Press.

Wertheimer, M. (1945). *Productive thinking.* New York: Harper & Row.

Whipple, G. M. (Ed.). (1930). *Report of the Society's committee on arithmetic.* (The twenty-ninth *Yearbook* of the National Society for the Study of Education.) Bloomington, IL: Public School Publishing Company.

Whitehurst, G. (2002). The Institute of Education Sciences: New wine, new bottles. Presentation at the 2002 annual meeting of the American Educational research Association, New Orleans, April 1–5, 2002. Retrieved April 1, 2004, from http://www.ed.gov/rschstat/research/pubs/ies.html.

Whitehurst, G. (2003). Evidence-based education. Powerpoint presentation dated June 9, 2003. Retrieved November 25, 2005, from http://www.ed.gov> on November 25, 2005.