

# METHOD

ALAN H. SCHOENFELD

*Elizabeth and Edward Conner Professor of Education*

*Graduate School of Education*

*University of California*

*Berkeley, CA 94720-1670, USA*

*Email: [alans@berkeley.edu](mailto:alans@berkeley.edu)*

Pre-Publication Draft Z: May 30, 2006

To appear in: F. Lester (Ed.), *Second handbook of research on mathematics teaching and learning*. New York: MacMillan

## METHOD

### Introduction

This chapter is concerned with research methods in mathematics education and, more broadly, with research methods in education writ large. As explained below, space constraints do not allow for the detailed consideration of individual methods, or even classes of methods. Hence I have chosen to address broad metatheoretical issues in much of the chapter, which is divided into three main parts. Part 1 provides an overview of the process of conducting and reflecting on empirical research. It examines the major phases of empirical research and some of the issues researchers must confront as they conduct their studies. A main thesis underlying the discussion in Part 1 is that there is a close relationship between theory and method. I describe the process of conducting empirical research and elaborate on how researchers' theoretical assumptions, whether tacit or explicit, shape what they choose to examine, what they see and represent in the data, and the conclusions they draw from them. Part 2 presents a framework for evaluating the quality of research. In it I argue that research must be judged by at least the following three criteria: trustworthiness, generality, and importance. A range of examples is given to elaborate on the issues discussed in Parts 1 and 2. In Part 3 I try to bring together the general arguments from the first two parts of the chapter by focusing methodologically on a topic of current interest and long-term importance. As this *Handbook* is being produced, there is great pressure on educational researchers and curriculum developers in the U.S. to employ randomized controlled trials as the primary if not sole means of evaluating educational interventions. In an attempt to move forward methodologically, I propose and discuss an educational analog of medical "clinical trials": the structured development and evaluation of instructional interventions. Part 3 offers a description of how the development and refinement of educational interventions might be conducted in meaningful ways, beginning with exploratory empirical/theoretical studies that reside squarely in "Pasteur's quadrant" (Stokes, 1997) and concluding with appropriately designed large-scale studies.

Before proceeding, I should comment about what this chapter is and is not. It is not a survey of research methods or (with the exception, in some sense, of Part 3) a "how to" guide to research. Such an approach would require a volume as large as this *Handbook* itself. Moreover, it would be largely redundant. There exist numerous handbooks of research methods in education, many weighing in at close to 1000 pages (see, e.g., Bruning & Kintz, 1987; Conrad & Serlin, 2005; Denzin & Lincoln, 2005; Green, Camilli, & Elmore, in press; Keeves, 1997; Kelley, & Lesh, 2000; LeCompte, Millroy & Preissle, 1992; Riley, 1990; Tashakkori & Teddlie, 2002). To give just one example of the extent of the methodological domain, the *Handbook of Complementary Methods in Education Research* (Green, Camilli & Elmore, 2006) contains chapters on 35 different research methods. The methods that begin with the letters C and D alone include: case studies: individual and multiple; cross-case analysis; curriculum assessment; data modeling: structural equation modeling; definition and analysis of data from videotape: some research procedures and their rationales; design experiments; developmental research: theory, method, design and statistical analysis; and discourse-in-use. It should be clear that even a cursory coverage of methods, much less a "how to," is beyond what can be done in this chapter.

What can and will be done is to take a bird's eye view of the terrain – to examine some overarching issues regarding the conduct of empirical research. It should be noted that from this perspective mathematics education is both special and not special. Mathematics education is special in that it is the focus of this *Handbook* and one of the best-mined fields of empirical research. All of the examples discussed in this chapter come from or serve to illuminate issues in mathematics education. At the same time, however, the issues addressed by these examples – What processes are involved in making sense of thinking, learning, and teaching? What are the attributes of high quality empirical research? How might one characterize a rigorous development and testing process for instructional interventions? – are general. The discussions in this chapter apply to all empirical research in education; indeed, to all empirical research.

**Part 1: On The Relationship Between Theory and Method; On Qualitative and Quantitative Methods; And a Framework for Examining Fundamental Issues Related to Empirical Inquiry.**

“There is no empirical method without speculative concepts and systems; and there is no speculative thinking whose concepts do not reveal, on closer investigation, the empirical material from which they stem.”

*Albert Einstein*

All empirical research is concerned with observation and interpretation. This is the case when one is crafting “rich, thick” descriptions (Geertz, 1975) of classrooms or of aboriginal cultures; it is also the case when one is conducting randomized controlled trials of rats running mazes after being subjected to different training regimes or of students taking mathematics assessments after being taught from different curricula. What may be less obvious, but is equally essential, is that all empirical research is concerned with and deeply grounded in (at times tacit but nevertheless strong) theoretical assumptions. Even the simplest observations or data gathering are conducted under the umbrella of either implicit or explicit theoretical assumptions, which shape the interpretation of the information that has been gathered. Failure to recognize this fact and to act appropriately on it can render research worthless or misleading.

In this opening part of this chapter I focus on issues of theory and method. First, I provide some examples to make the point that theory and method are deeply intertwined – that, as the quotation from Einstein attests, there are no data without theory and there is no theory without data. Then I proceed to put some flesh on the bare bones of this assertion. I offer a framework for conducting and examining empirical research. Readers are taken on two “tours” of this framework, one describing an example of qualitative research and one describing an example of quantitative research. A main point of the discussions is to show that divisions between the two types of research are artificial – that the same theoretical and empirical concerns apply to both.

*On Framing Questions, Data Gathering, and Questions of Values*

From the onset of a study, the questions that one chooses to ask and the data that one chooses to gather have a fundamental impact on the conclusions that can be drawn.

Lurking behind the framing of any study is the question of what is valued by the investigators, and what is privileged in the inquiry.

For example, a recurrent issue in college level mathematics is typically posed as follows: “Is there evidence that small classes (e.g., recitation sections with thirty or fewer students) are more effective than large lecture classes?” What must be understood is that the way this question is operationalized and the choice of evidence that will be used to inform a decision are consequential.

One way to judge course effectiveness is to examine student scores on a uniform end-of-term examination. For reasons of efficiency, students in large lecture classes are often tested using skills-oriented multiple choice tests. Thus, one might decide to give such tests to students in both small and large calculus classes, and look for differences in scores.<sup>1</sup> It might well be the case that on such a test there would be no statistically significant differences between the scores of students in large and small classes. On the basis of this evidence, the two forms of instruction could be judged equivalent. Once that judgment has been made, cost might be used as the deciding factor. The institution might opt to offer lecture classes with large enrollments.

An alternative way to evaluate course effectiveness is to look at the percentage of students in each instructional format who enroll in subsequent mathematics courses or who become mathematics majors. With that form of evaluation, small classes might produce better results. On the basis of such evidence, the institution might decide (cost factors permitting) to offer classes with small enrollments.

The point of this example is that both test scores and subsequent enrollment rates are legitimate measures of the outcomes of instruction. Each can be quantified objectively and used as the justification for policy decisions. Yet, the two measures might lead to different conclusions. A decision to use one measure or the other, or a combination of both, is a reflection of one’s values – a reflection of what one considers to be important about the students’ experience. In this sense, even simple quantitative data gathering and analysis are value-laden. The same is the case for qualitative analyses. Historians, for example, will decide that certain pieces of evidence in the historical record are relevant to their framing of an historical issue while others are not. These acts of selection/rejection are consequential for the subsequent representation and analysis of those data.<sup>2</sup>

### *On the Relationship Between Theory and Data*

In recent years “evidence-based medicine” (see, e.g., the Cochrane Collaboration at <http://www.cochrane.org/index0.htm>) has been advocated by some, notably by federal administration figures such as Grover Whitehurst, director of the U. S. Department of Education’s Institute for Education Sciences, as a model for how to conduct empirical research in education (see, e.g., Whitehurst, 2003). For this reason I have selected as

<sup>1</sup> How well a skills-oriented test might actually reflect what a group of students has learned, and what conclusions can be drawn from such using such tests, are serious matters. Those issues are considered in the discussion of Ridgway, Crust, Burkhardt, Wilcox, Fisher and Foster (2000) later in this chapter.

<sup>2</sup> N.B. Historians, and social scientists in general, often make their cases via narrative. One must understand that narrative is a form of representation; my comments about representations apply to narrative work as well.

cases in point for this discussion of the relationship between theory and data some uses of the experimental paradigm in medical research.<sup>3</sup> Direct connections to research in mathematics education will be drawn after the examples have been presented.

Consider as an example the use of the “male norm” in clinical studies (Muldoon, Manuck, & Matthews, 1990; National Research Council, 1994; World Health Organization, 1998; Wysowski, Kennedy, & Gross, 1990), in which the results of male-only studies have been assumed to apply to both men and women. The March 2005 issue of the *New England Journal of Medicine* reported the results of a 10-year study of women’s use of low-dose aspirin to combat heart disease. Among the findings are the following. In contrast to the situation with men, for whom taking low-dose aspirin on a daily basis has consistently been shown to lower the likelihood of heart attacks, taking a low daily dose of aspirin did not, overall, reduce the likelihood of a first heart attack or death from cardiovascular disease for women. However, there were age-specific results: Aspirin did substantially reduce the likelihood of heart attacks in women over the age of 65. Similarly, recent medical research indicates that there are differential risks of diabetes for different subpopulations of the general population.

There is a sampling point here: assuming that the results of a study (no matter how well executed) that is conducted on a subpopulation will apply to the population as a whole is not necessarily warranted. Selecting an appropriate sample is a subtle art, and unexamined assumptions may skew a sample badly. Conversely, studies that average results over an entire population may fail to reveal important information about specific sub-populations – that is, averages may mask important effects. (See, e.g., Siegler, 1987, and the discussion of Bhattacharjee, 2005, below.)

This example also makes an equally important point regarding the researchers’ underlying conceptual models. When “male norm” studies were paradigmatic, the assumption was that a random selection of males was a random selection of people – that gender didn’t matter. That is, the experimenters did not consider gender to be a relevant variable in their experiments. This failure to conceptualize gender as a variable rendered the studies of questionable value.

In sum: Whether it is tacit or explicit, one’s conceptual model of a situation, including one’s view of what counts as a relevant variable in that situation, shapes data-gathering – and it shapes the nature of the conclusions that can be drawn from the data that are gathered. As will be discussed later in this chapter, issues such as the characteristics of the student population (e.g., what percentage of students are second-language learners?) or of the environment (e.g., is the school capable of implementing a curriculum as intended?) can be fundamental factors shaping what takes place in a learning environment. Whether and how those factors are taken into account in formulating a study and gathering data for it will shape how that study’s findings can be interpreted.

A second issue, touched on in the class size example discussed above, concerns the experimenter’s selection of outcomes (dependent variables) and the selection of

---

<sup>3</sup> Throughout this chapter I discuss examples of significant current interest such as controversies over randomized controlled trials as the “gold standard” for educational research. In doing so I am attempting to achieve simultaneously the dual goal of addressing enduring points of concern and clarifying current issues.

measures to document those outcomes. To give a medical example: By the 1990s hormone replacement therapy (HRT) had become a commonly recommended treatment for some of the symptoms of menopause. When subsequent research examined an expanded set of outcomes such as the incidence of heart disease, breast cancer, and strokes, the value of HRT was called into question (see Medline Plus, 2005, for an overview). Delayed or unexpected consequences are also an issue. The devastating impact of thalidomide was not discovered until some years after the drug had been in common use.

It may seem quite a leap to compare the results of such medical studies with the results of educational interventions. However, there are direct analogues. Like medical interventions, educational interventions can have unintended and often long-term consequences. For example, a body of research in the 1970s and 1980s, which included the qualitative documentation of classroom interactions and results, documented the results of students' school mathematics experiences. These were summarized by Lampert (1990) as follows:

Commonly, mathematics is associated with certainty; knowing it, with being able to get the right answer, quickly (Ball, 1988; Schoenfeld, 1985b; Stodolsky, 1985). These cultural assumptions are shaped by school experience, in which doing mathematics means following the rules laid down by the teacher; knowing mathematics means remembering and applying the correct rule when the teacher asks a question; and mathematical truth is determined when the answer is ratified by the teacher. Beliefs about how to do mathematics and what it means to know it in school are acquired through years of watching, listening, and practicing. (p. 32)

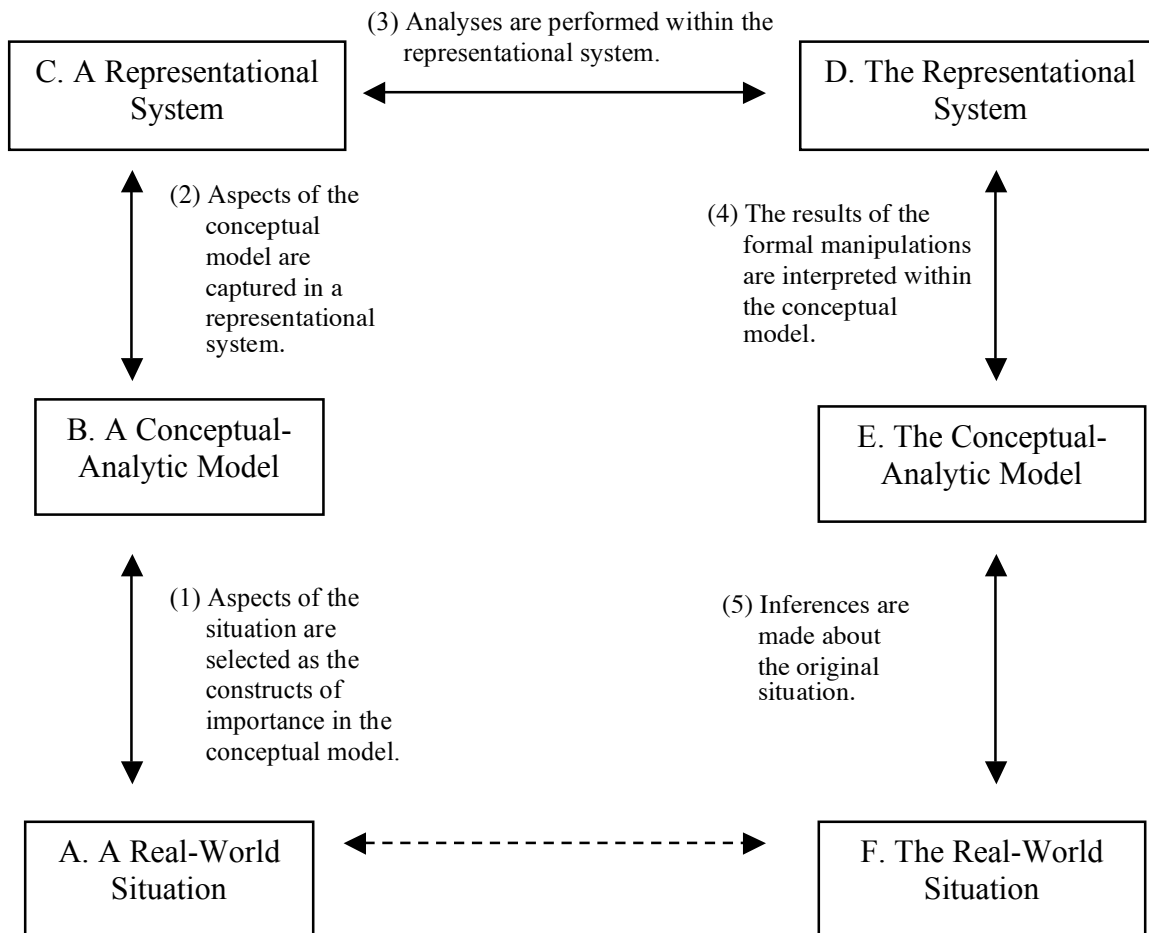
Let me reframe this summary in terms of contemporary discussions. As Lampert indicated, years of learning mathematics passively result in a population that tends to be mathematically passive. That population may be able, on demand, to perform some mathematical procedures – but it tends not to possess conceptual understanding, strategic competency, or productive mathematical dispositions. If the measures and descriptions of educational outcomes that are employed in empirical research fail to take into account such centrally important classes of outcomes (e.g., conceptual understanding as well as procedural competency; the ability to apply one's knowledge to novel concepts; problem-solving ability; beliefs and dispositions; drop-out rates), then researchers, teachers, and policymakers who wish to make judgments on the basis of those outcomes are potentially misinformed about the likely consequences of their decisions.

#### *A Framework for Conducting and Examining Empirical Research*

The preceding metatheoretical comments frame what is to come in this section. In what follows I set forth a framework for conceptualizing empirical work, whether that research is qualitative or quantitative, in *any* field. Figure 1 (modified from Schoenfeld, 2002, with permission) offers a framework within which to consider issues of method. After an introductory caveat, I briefly introduce the framework. Then I work through it in some detail.

*Caveat.* The discussion of Figure 1 proceeds in an ostensibly straightforward manner, from the “beginning” of the research process (conception and formulation of problems) to its “end” (drawing conclusions). However, the linear nature of the

exposition and the representation in Figure 1 belie the complexity of the process, which is decidedly non-linear as it plays out in practice. Research proceeds in cycles, in which one considers and then reconsiders every aspect of the process. Even within cycles, insights (including those caused by failure or chance observation) may cause a reformulation of underlying perspective, or of what are considered salient phenomena; they may result in new representations, alternative data gathering or new ways of thinking about data that have already been gathered; and new conclusions. Specifically, Figure 1 is not to be taken as a linear prescription for research.



*Figure 1.* A schematic representation of the process of conducting empirical research.

In simplest terms, empirical research is concerned with making observations of and drawing conclusions about some “real world” situation. Data are gathered and interpreted, and conclusions are drawn. That process is represented by the dotted line from Box A to Box F in Figure 1. The conclusions drawn are sometimes just about the situation itself (“I observed the following...”), but more typically they are drawn with intimations of generality (“What happened here is likely to be the case in circumstances that resemble those described here.”) and importance (“This information should shape the ways we think about X, Y, and Z.”). The main purpose of Figure 1 is to indicate that the pathway from observations to conclusions is not as simple as it might seem.

In line with Figure 1, I claim that all empirical research involves the following processes:

- conceptualization, in which the situation to be analyzed is seen and understood in certain (often consequential) ways;
- the creation, use, or refinement of a conceptual-analytic framework or model, in which specific aspects of the situation are singled out for attention (and, typically, relationships among them are hypothesized);
- the creation, use, or refinement of a representational/analytic system, in which aspects of the situation singled out for attention are selected, represented and analyzed;
- the interpretation of the analysis within the conceptual-analytic framework or model; and
- attributions and interpretations from the preceding analytic process to the situation of interest (and possibly beyond).

To illustrate the main points above I consider at some length one primarily qualitative example and one primarily quantitative example.

*A first example.* As a first qualitative example I discuss the decision I made, nearly 30 years ago, to explore aspects of students' metacognitive behavior during problem solving. (Extensive detail regarding this work can be found in my 1985 book *Mathematical Problem Solving*.) The starting place for this work seemed simple. I brought students (either by themselves or in pairs) into a room near my office (my "laboratory") and asked them to solve a series of problems out loud. I was in the vicinity while they worked on the problems, and I occasionally intervened if a long time had passed without their saying something audible. I videotaped the students' solution attempts and saved their written work.

The primary sources of data for analysis were the videotapes I made of their problem-solving attempts and the written work they produced while working the problems. On the basis of those data I drew inferences about the students' decision making during problem solving and its impact on their success or failure at problem solving. I also drew inferences about the frequency and import of the students' "executive decision making" in general.

To illustrate the issues involved, I start with Box A at the lower left of Figure 1, and make a circuit of the figure by following the arrows up, around, and down to Box F. To begin, it should be clear that I was making a fair number of assumptions about the "real world situation" examined here – students solving problems in the laboratory. Two major assumptions were that (a) the students' problem solving behavior in the laboratory bore some relation to their problem solving behavior in other contexts; and (b) the students' overt actions bore some relation to their internal cognitive processes.

Both of these assumptions were and are controversial to some degree. Regarding (a), for example, over the years some researchers have questioned the value of laboratory studies, saying that the artificial behavior induced in the laboratory renders laboratory studies of little or no value in understanding the kinds of interactions that take place

amidst (for example) the blooming complexity of the classroom. Regarding (b), for quite some time there have been controversies over the role of verbal reports as data. Retrospective reports of thought processes were roundly discredited in the early years of the 20th century, and for some years *any* reports of thought processes were deemed illegitimate. (Indeed, behaviorists banished the notion of thought processes from “scientific” explanations of human behavior.) In the 1980s Nobel prize winner Herbert A. Simon and colleague K. Anders Ericsson wrote a review (Ericsson & Simon, 1980) for *Psychological Review* and then a book (Ericsson & Simon, 1984) entitled *Verbal Reports As Data*, trying to make the case that although post hoc reports of thought processes could not be taken as veridical, “on the spot” verbalizations of what one was doing could be taken as data suggestive of the individuals’ thought processes.

One could say a great deal more about assumptions (a) and (b) – teasing out what “some relation” means in each of them is a nontrivial exercise! What matters here is something simpler. Wherever one comes down with regard to assumptions (a) and (b), the fact is that they *are* assumptions, and one’s stance toward them shapes how one considers the data gathered. What should be clear is that a form of naïve realism – that the videotapes and written record *directly* capture (some of) what people were thinking as they worked on the problems – is not warranted. Equally clear is that I began my work with specific assumptions about what “out loud” problem-solving protocols could reveal; I entered into the work with a set of underlying assumptions about the nature of cognition that framed the way I saw what was in the tapes. Someone who was not a cognitive scientist, or whose orientation to cognition was different, would not look for or see the same things.

When I began examining the videotapes, I knew there was *something* important about students’ decision making during problem solving – something that was a factor in success or failure – but I did not know what it might be. My earlier work had focused on teaching an explicit decision making strategy, to help students use their problem-solving knowledge effectively. Now I was looking at videotapes made before the instruction, trying to identify causes of success or failure. I was looking at the tapes “from scratch” in part because the fine-grained coding schemes I had found in the literature had not seemed informative.

My research assistants and I watched a fair number of tapes, trying to figure out how to capture events of importance in a coding scheme. We started in a somewhat systematic way, looking for what we called “reasons to stop the tapes.” These occurred at places in the videotapes where we saw students acting in ways that seemed to bear on the success or failure of their problem solving attempts. We made a list of such events and composed for each event a series of questions designed to trace its impact on the problem solution. This was a prototype analytic scheme. And after polishing it a bit I asked my students to try to analyze the data using it.

When we reconvened, my research assistants were unhappy. They said that the scheme we had developed was impossible to use. Far too many of our questions, which had seemed to make sense when we looked at one tape, seemed irrelevant on another. Our system had so many reasons to stop a tape, and so many unanswerable or irrelevant questions when we did, that whatever was truly important about the problem-solving episode was lost among the huge collection of questions and answers.

Confronted with this failure, I decided to begin again. I chose to look at an “interesting” tape – a tape in which it seemed that the students “should have” solved the problem but did not. My assistants and I tossed the coding scheme aside and looked at the tape afresh. As we did, I noticed one particular decision that the students in the videotape had made. They had chosen, without much deliberation, to perform a particular computation. As the solution unfolded, they spent a great deal of time on the computation, which I realized would not help them to solve the problem. As I watched them persevere in the computation, things clicked. That single decision to perform the computation, unless reversed, could result in the expenditure of so much time and energy in an unprofitable direction that the students were essentially guaranteed to fail to solve the problem.

I had the feeling I was on the trail of something important. My assistants and I looked at more tapes, this time searching for consequential “make-or-break” decisions. It turned out that these were of two kinds: paths wrongly taken and opportunities missed. These make-or-break decisions were consequential in more than half of our tapes. With this understanding, we had a new perspective on what counts as a major factor in problem solving. This new conceptual/analytic perspective oriented us differently toward the tapes and changed our subsequent data analyses. At this point, with a conceptual model in place, we were in Box B of Figure 1.

[Before proceeding, I must stress that not every study involves a new conceptual model; most studies involve the use or refinement of well-established conceptual models. The point of this particular example is that any conceptual model highlights some things and obscures or ignores others; it takes some things into account and does not consider others. For example, my analyses of the videotapes of students solving problems did not, at that point, include a focus on issues of affect or belief. They did not include the detailed examination of student knowledge or knowledge organization, save for the fact that I had been careful to have the students work problems for which I had evidence that they possessed adequate knowledge to obtain a solution. (It is of little theoretical interest when a student fails to solve a problem simply because he or she lacks the knowledge that is essential to solve it.) Hence, as I was examining the problem-solving tapes, I was viewing them through a particular theoretical lens, one that focused on the impact of a particular kind of decision making. The videotapes might well have supported different kinds of analyses, but other aspects of the students’ solutions were not to be seen in our analyses (and, equally important, ceased to be salient to us as we analyzed the tapes). I also note that this example demonstrates the dialectic between representational/analytic schemes and conceptual frameworks, thus illustrating the non-linear character of Figure 1.]

Once my research assistants and I had a first-order conceptual-analytic framework, we needed a representational scheme to capture and analyze our data. In simplest terms, we decided to parse problem-solving sessions into major chunks called “episodes,” periods of consistent goal-oriented activity on the part of the problem solver. The notion of an episode was a useful device for identifying the loci of consequential decisions. The places where the direction of a solution changed were the natural boundaries between episodes, and they were often the sites of consequential decisions. It also turned out that, at a gross level, there were relatively few kinds of episodes: reading

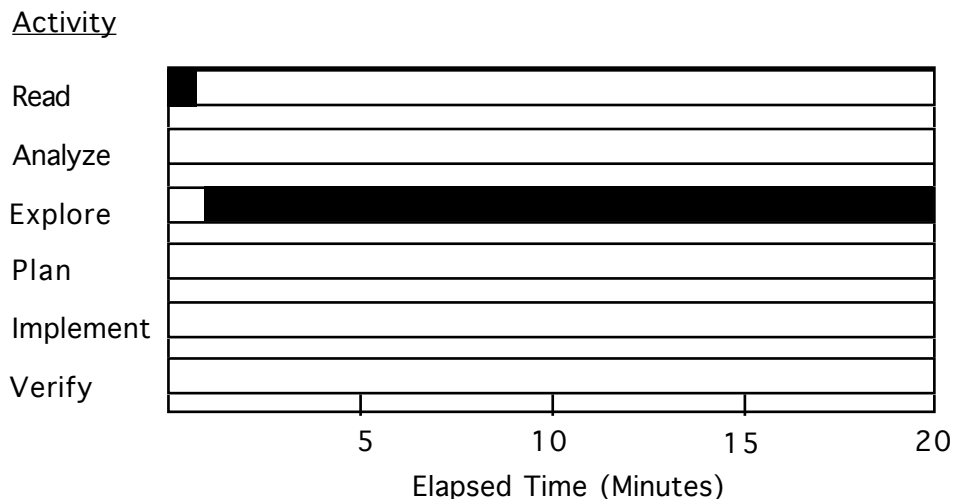
the problem, working in a structured way to analyze the problem, planning, implementing a plan, and working in a somewhat random or ill-thought-out way (“exploration”).

Over time we refined this representational scheme, which was later supplemented by a more compact and informative time-line representation suggested by Don Woods (Figure 2, below, is an example). With a representational scheme in place, we were able to code the data. Things were straightforward. The idea was to represent the contents of a videotape (typically a 20-minute problem-solving session) by an episode diagram, which identified and labeled the episodes and the consequential decisions in a problem session.

At this point we were comfortably ensconced in Box C, working within a particular representational system. It is important to observe that the representational system reified our conceptual model. Events that did not fit into the representational scheme were not captured in the representation, and thus were not fodder for data analysis.

My research assistants and I worked together on a series of tapes, developing a communal understanding of the meanings and types of episodes, and of consequential decision making (including the failure to act on a relevant piece of information). We then coded some tapes independently. Our codings matched more than 90% of the time.

With the consistency of coding established, we were working within Box D – performing analysis with and within the representational system. Coding the sessions was still a matter of interpretation, but with practice it became a relatively straightforward task, as indicated by the high interrater reliability. The hard work had been done in the conceptualization of the scheme. Once the tapes were represented in the coding scheme, data analysis was simply a matter of counting. More than half of the students’ problem-solving attempts were represented by the following schematic diagram: an episode of reading the problem followed by an episode of exploration (and failure). That is, the bar graph in Figure 2 represented more than half of the problem-solving sessions we coded.



*Figure 2.* A timeline representation of a typical student solution attempt.

These data had a straightforward interpretation. More than half the time, the students – who had the knowledge required to solve the given problems – failed to solve

the problems because of their poor choice of initial solution direction. As long as the students lacked effective mechanisms for reflecting on and undoing those initial decisions, they did not have the opportunity to use what they did know to solve the problems.

This interpretation placed us in Box E, and the extrapolation to Box F (and beyond) was also straightforward. We recognized, of course, that factors other than monitoring and self-regulation affected problem solving success – knowledge and problem-solving strategies among them. In our work, we had controlled for these by giving the students problems that they had the subject matter knowledge to solve; thus we saw the devastating effects of poor metacognitive decision making. We asserted that ineffective monitoring and self-regulation were significant causes of problem-solving failure, both in the laboratory and beyond it.

As noted, there were many assumptions made here, in the attribution of causality both in laboratory problem solving and in the extrapolation to more general problem solving. At this point, given the small number of students we had videotaped, our ideas about the importance of metacognitive decision making were suggestive but not yet validated. Further studies led to substantiation of those ideas. The study of accomplished problem solvers (in the laboratory) documented the ways that effective monitoring and self-regulation can be a productive force in problem solving. Subsequent research showed the robustness of the phenomena of monitoring and self-regulation, and their importance. (See, e.g., Brown, 1987; deCorte, Green, & Verschaffel, 1996; Lester, 1994.)

*A second example.* I now examine the issues raised by a prototypical quantitative study (Bhattacharjee, 2005). Again, I focus on the kinds of choices that are made at various points in the process described in Figure 1 and the impact they have on the conclusions that can be drawn. Consider the task of determining whether students learn more from Curriculum X or from Curriculum Y. As many people see it, this is as straightforward as you can get: All one has to do is perform some sort of randomized controlled trial in which half the student population is randomly assigned each treatment, and then see what differences emerge on an appropriate outcome measure. Would that life were so simple.

The complexities begin in the real-world context, Box A. Even before articulating a conceptual model, there are choices at the pragmatic level. Does one randomize the instructional treatment at the student level (with two adjacent students perhaps studying different curricular materials)? At the classroom level? At the school level? Such considerations are often driven by practicalities. But then, when one goes to the conceptual level (Box B), choices get much more complex – and consequential.

At least two kinds of conceptual issues are fundamental in shaping how one does research on curricular outcomes. The first is methodological, the second subject-matter related. A fundamental conceptual issue related to curriculum implementation is what one considers an “implemented curriculum” to be. One perspective is as follows.

*Perspective 1.* A curriculum is the set of instructional materials and preparation to use them that teachers are given. Whatever the teachers do with those materials in the classroom is the “implemented curriculum.”

In this case, what counts as the simplest measure of the curriculum's effectiveness is the average performance of all those students who were in classrooms where that curriculum was used.

Another perspective is as follows.

*Perspective 2.* There is a strong degree of interaction between curriculum and context. Given different contexts or different degrees of support, there may be more or less fidelity of curriculum implementation. "Degree of fidelity of implementation" (in conformity with the intention of the designers) matters and should be taken into account in analyses of curriculum impact.

Questions of interest to people with this orientation include the following. What kinds and levels of support are necessary, in what kinds of contexts, to guarantee some level of fidelity of implementation for a particular curriculum? What is the curricular impact (measured in terms of student outcomes) when there is some degree of curriculum fidelity? When one is considering a choice between two curricula, what kinds of outcomes can one expect for each, given the resources that one has to put into implementing them?

This distinction in framing has significant consequences. Here is an example, taken from a recent issue of *Science* (Bhattachargee, 2005). In a randomized trial in a school district, three schools used Curriculum X and three used Curriculum Y. The schools were roughly matched on demographics. When one looked at overall curriculum implementation – that is, the complete sets of scores from students who had worked through Curriculum X and Curriculum Y – no statistically significant differences between outcomes were found. Is one to conclude, then, that the two curricula are equally good?

The answer depends on one's conceptual model. For those who adhere to perspective 1 as described above, the situation is straightforward. Given that the schools were randomly assigned to the treatments and the data showed no differences, it follows (within perspective 1) that the curricula are equally effective. But for those who hold perspective 2, there might be a world of difference between the two curricula.

It turns out that of the three schools that used Curriculum X, one school embraced the curriculum and implemented it in a way consistent with the designers' intentions. Students at that school outperformed those who used Curriculum Y. At a second school the implementation of Curriculum X was uneven. There, scores were not statistically different from overall scores on Curriculum Y. In a third school Curriculum X was poorly implemented, and students did poorly in comparison to Curriculum Y.

"On average" the two curricula were equally effective. The averages are uninformative, however. Another way to look at the data is as follows. When Curriculum X is implemented as intended, outcomes are superior to outcomes from Curriculum Y. Under those conditions, Curriculum X is preferable. But when Curriculum X is not implemented effectively, students would do better with Curriculum Y. Hence instructional leadership should assess the capacity of the staff at each site to implement Curriculum X – either now or with professional development – and decide on that basis whether to use it at that site. From perspective 2, then, the decision as to whether to use

Curriculum X or Curriculum Y is context-dependent, depending on the school staff's current or potential capacity to implement either curriculum with some degree of fidelity. Note that this is a very different kind of conclusion than the kind of conclusion drawn by those with the "curriculum is context-independent" perspective.

Here is a relevant analogy. Suppose there are two surgical treatments for a particular condition. Treatment A returns patients to full functioning *if* they undergo a full regimen of physical therapy for a year, but the results are unsatisfactory if a patient does not. Treatment B is reasonably but not completely effective, regardless of whether the patient undergoes physical therapy. Suppose that, on average, not that many people follow through on physical therapy. On average, then, people do slightly better with Treatment B than with Treatment A.

Put yourself in the position of a patient who is facing surgery for that condition. Would you want your doctor to recommend Treatment B on the basis of the statistical average? Or would you rather have the doctor explain that Treatment A might be an option for you, but only if you commit yourself to a serious regimen of physical therapy afterward? Both statements represent legitimate interpretations of the data, within the frames of particular conceptual models. Those models make a difference. As a patient, I would much rather be offered the second choice. There is no reason to settle for the statistical average if there are reliable ways to beat that average. (One should settle for it, however, if one does not have the wherewithal to follow through with physical therapy.)

To return to the curricular example: one's conception of what is meant by "curriculum implementation" has tremendous implications for the ways that findings are reported and interpreted. One can report on data of the type discussed by Bhattacharjee (2005) either by saying

- (a) "There were no significant differences between Curriculum X and Curriculum Y" or
- (b) "Curriculum X is superior to Curriculum Y under certain well-specified conditions; Curriculum X and Curriculum Y produce equivalent test scores under a different set of well-specified conditions; and Curriculum Y is superior to Curriculum X under yet another set of well-specified conditions."

The possibilities for *acting* on the information in (a) and (b) differ substantially.<sup>4</sup> I now consider conceptual models related to subject matter.

Just what does it mean to know (to have learned) some particular body of mathematics? This is not only a philosophical issue, but a practical one as well: Different conceptual models of mathematical understanding lie at the heart of the "math wars" (see Schoenfeld, 2004). One point of view, which underlies much of the "traditional" curricula and standardized assessments, is that knowledge of mathematics consists of the mastery of a body of facts, procedures, and concepts. A more current perspective, grounded in contemporary research, is that mathematical knowledge is more complex. The "cognitive revolution" (see, e.g., Gardner, 1985) produced a fundamental epistemological shift regarding the nature of mathematical understanding. Aspects of mathematical competency are now seen to include not only the knowledge base, but also

---

<sup>4</sup> This idea is not new: see, e.g., Brownell, 1947.

the ability to implement problem-solving strategies, to be able to use what one knows effectively and efficiently, and more (deCorte, Greer, & Verschaffel, 1996; Lester, 1994; Schoenfeld, 1985a, 1985b, 1992). In elementary arithmetic, for example, the National Research Council volume *Adding It Up* (2001) described five interwoven strands of mathematical proficiency:

- *conceptual understanding* – comprehension of mathematical concepts, operations, and relations
- *procedural fluency* – skill in carrying out procedures flexibly, accurately, efficiently, and appropriately
- *strategic competence* – ability to formulate, represent, and solve mathematical problems
- *adaptive reasoning* – capacity for logical thought, reflection, explanation, and justification
- *productive disposition* – habitual inclination to see mathematics as sensible, useful and worthwhile, coupled with a belief in diligence and one’s own efficacy. (p. 5)

Fine-grained analyses of proficiency tend to be aligned with the content and process delineations found in the National Council of Teachers of Mathematics’ (NCTM, 2000) *Principles and Standards for School Mathematics*:

*Content*: Number and Operations; Algebra; Geometry; Measurement; Data Analysis and Probability;

*Process*: Problem Solving; Reasoning and Proof; Making Connections; Oral and Written Communication; Uses of Mathematical Representation.

These views of proficiency extend far beyond what is captured by traditional content-oriented conceptual frameworks.

In the experimental paradigm, one’s view of domain competency is typically instantiated in the tests that are used as outcome measures. What view of mathematical proficiency one holds, and how that view is instantiated in the outcome measures one uses for educational interventions, can make a tremendous difference.

The issues at stake are as follows. Traditional assessments tend to focus on procedural competency, while assessments grounded in broad sets of standards such as NCTM’s *Curriculum and Evaluation Standards* (1989) or *Principles and Standards* (2000) include procedural (skills) components but also assess conceptual understanding and problem solving. In a rough sense, the traditional assessments can be seen as addressing a subset of content of the more comprehensive standards-based assessments. Hence a choice of one assessment instead of another represents a value choice – an indication of which aspects of mathematical competency will be privileged when students are declared to be proficient on the basis of test scores. As the following example shows, these choices are consequential.

Ridgway, Crust, Burkhardt, Wilcox, Fisher, and Foster (2000) compared students’ performance at Grades 3, 5, and 7 on two examinations. The first was a

standardized high-stakes, skills-oriented test – California’s STAR test, primarily the SAT-9 examination. The second was the Balanced Assessment test produced by the Mathematics Assessment Resource Service, known as MARS. The MARS tests cover a broad range of skills, concepts, and problem solving. For purposes of simplicity in what follows, scores on both tests are collapsed into two simple categories. Student who took both tests are reported below as being either “proficient” or “not proficient” as indicated by their scores on each of the examinations. More than 16,000 students took both tests. The score distribution is given in Table 1.

MARS	SAT-9	
	Not Proficient	Proficient
Grade 3 ( $N = 6136$ )		
Not proficient	27%	21%
Proficient	6%	46%
Grade 5 ( $N = 5247$ )		
Not proficient	28%	18%
Proficient	5%	49%
Grade 7 ( $N = 5037$ )		
Not proficient	32%	28%
Proficient	2%	38%

*Table 1.* Comparison of Students’ Performance on Two Examinations

Unsurprisingly, there is a substantial overlap in test performance: Overall 73%, 77%, and 70% of the students at Grades 3, 5, and 7, respectively, either passed both tests or failed both tests. The interesting statistics, however, concern the students who were rated as proficient on one test but not the other.

For each grade, consider the row of Table 1 that reports the SAT-9 scores for those students rated “proficient” on the MARS test. At Grades 3, 5, and 7 respectively, 88%, 91%, and 95% of those students were rated proficient on the SAT-9. Thus being rated proficient on the MARS test yields a very high probability of being rated proficient on the SAT-9. That is: being declared proficient on the MARS exam virtually assures having the procedural skills required for the SAT-9.

The converse is not true. Consider the final column of Table 1, which indicates the MARS ratings of the students who were rated proficient on the SAT-9. Approximately 31% of the third graders, 27% of the fifth graders, and 42% of the fifth graders who were declared proficient by the SAT-9 were declared not proficient on the MARS exam. That is, possessing procedural fluency as certified by the SAT-9 is clearly *not* a guarantee that the student will possess conceptual understanding or problem-solving skills, as measured by the MARS test. Indeed, the students who were declared proficient on the SAT-9 but

not the MARS test – roughly 1/3 of those declared proficient on the SAT-9 – can be seen as false positives, who have inappropriately been deemed proficient on the basis of a narrow, skills-oriented examination.

Once an assessment has been given, the die has been cast in terms of data collection. One is now in Box C in Figure 1, where there exist standard techniques for scoring tests and representing test data. The pathway from Box C to Box D in Figure 1 is relatively straightforward, as are analyses within Box D. This, after all, is the province of standard statistical analysis. However, *interpretation* – the pathway to Boxes E and F – is anything but straightforward, for it depends on the conceptual models being employed.

There is suggestive, though hardly definitive, evidence (see, e.g., Senk & Thompson, 2003) that nearly all of the National Science Foundation-supported standards-based curricula have the following property. When the test scores of students who have studied from the NSF-supported curricula are compared with test scores of students who have studied from more traditional skills-oriented curricula, there tend to be no statistically significant differences between the two groups in performance on skills-oriented tests (or the skills components of broader tests). However, there tend to be large and significant differences favoring the students from the NSF-supported curricula on measures of conceptual understanding and problem solving. Thus, if appropriately broad assessments are used, comparison studies will tend to produce statistically significant differences favoring the performance of students in these standards-based curricula over the performance of students from more traditional comparison curricula. However, if skills-oriented assessments are used, no significant differences will be found. Hence at the curriculum level, the use of measures that focus on skills can result in curricular false negatives – the tests will fail to show the real differences that exist.

The fundamental point to be taken from the preceding discussion is that the specific contents of any given assessment matter a great deal. One can draw meaningful conclusions about the relative efficacy of two curricula on the basis of a particular assessment only when one knows what the assessment really assesses (that is, when a content analysis of that assessment has been done). Without a content analysis, it is impossible to interpret a finding of “no significant differences.” Such a finding might occur because both curricula are equally effective. Or, it might occur because an inappropriately narrow assessment failed to pick up what are indeed significant differences in impact. For this reason, a report of a randomized controlled trial that does not contain a content analysis of the assessment employed is of no value. Indeed, the conclusions drawn from it may be false or misleading.

Ironically, this is the mistake made by the nation’s most ambitious attempt to provide information about curricular effectiveness, the What Works Clearinghouse (WWC). WWC (<http://www.whatworks.ed.gov/>) does not conduct research itself. Rather, it was created to review and report findings from the literature. WWC searches the literature for studies that meet stringent methodological criteria. Studies that qualify for vetting by WWC must be of one of the following three types: randomized controlled trials, quasi-experiments that use equating procedures, or studies that use regression discontinuity designs. These are vetted for technical proficiency and empirical flaws. Only studies that make it through WWC’s methodological filter are reported.

WWC committed the fundamental error identified above in reporting one of the few studies that did make it through its methodological filter. In a report (What Works Clearinghouse, 2004), WWC gave part of the statistical analyses in the study it examined (a quasi-experimental design with matching reported in 2001 by C. Kerstyn) full marks. Here is what WWC (2004) said about its choice of that part of the study:

The fifth outcome is the Florida Comprehensive Assessment Test (FCAT), which was administered in February 2001. The author does not present the reliability information for this test; however, this information is available in a technical report written by the Florida Department of Education (2002). This WWC Study Report focuses only on the FCAT measures, because this assessment was taken by all students and is the only assessment with independently documented reliability and validity information.

Note that reliability and validity are psychometric properties of an assessment: They do not provide a characterization of the actual content of the examination. Neither Kerstyn nor WWC conducted content analyses of the FCAT exam. For all one knows, it could be as narrow as the SAT-9 examination discussed by Ridgway et al. (2000). The Kerstyn study reported “no significant differences” – but why? Was it because there were none, or because the narrowness of the measure used failed to reveal a significant difference that actually existed? Because of the lack of information provided by WWC, it is impossible to know. Given that WWC failed to conduct a content analysis of the FCAT, the findings reported in the WWC report are at best worthless and at worst misleading. In addition, WWC’s unwillingness to conduct content analyses of the measures used in the randomized controlled trials of mathematics studies makes it impossible for WWC to achieve its core mission. WWC was created with the intention of conducting meta-analyses of the literature – to sort out through analytical means the impact of various curricula. Properly conducted, the analyses and meta-analyses are intended to reveal information such as the following: “Curriculum X tends to be strong on procedural skills and on conceptual understanding, but not especially strong on problem solving. Students tend to do well on tests of geometry, measurement, and number, but they do less well on tests of algebra and data analysis.” Given that WWC has refused to conduct content analyses<sup>5</sup>, WWC can offer no insights of this type. Once again, what is attended to, both in conceptual models and in assessments, is highly consequential.

In sum, although one must be proficient in the application of quantitative and qualitative methods on their own (specifically, the pathway from Box C to Box D in Figure 1), such proficiency is no guarantee that the interpretation of the results will be meaningful or useful. A meaningful report must respect all of the pathways from Box A to Box F in Figure 1.

### *Discussion*

In this section I have focused on some fundamental issues of theory and method. First, I argued that theory and method are deeply intertwined. Every empirical act of

---

<sup>5</sup> I served as the Senior Content Advisor for WWC’s mathematics studies (at first for middle school mathematics, then for all mathematics reports) from WWC’s beginnings. I resigned in early 2005 when WWC refused to correct the flaws identified above and reneged on a commitment to publish an article in which I had discussed such issues. For details see Schoenfeld (2006).

representation, analysis, and interpretation is done in the context of a (sometimes explicit, sometimes implicit) conceptual and theoretical model. The character of such models shapes the conclusions that are produced by subsequent analysis and interpretation. Second, I have presented a framework (Figure 1) that highlights major aspects of empirical research including conceptualization, representation, analysis, and interpretation. I remind the reader that although the figure and the linearity of prose as a medium may suggest that the process is linear, it is not: the process is cyclical, and there can be substantial give-and-take between all of the aspects of research reflected in Figure 1 during each cycle of research. The extensive discussion of Figure 1 highlighted the complexity of the process and the ways in which conceptual models can affect what one captures in data and how those data are interpreted. Third, I deliberately chose to work through one prototypically qualitative and one prototypically quantitative example to indicate that the fundamental issues of focus, data gathering, data analysis, and interpretation of findings are the same whether one is conducting qualitative or quantitative research.<sup>6</sup> The serious question to be considered is not, “is this research of one type or another” but “what assumptions are being made, and how strong is the warrant for the claims being made?”

Finally, I want to point to the fact that the framework outlined in Figure 1 can be used reflectively, both as one conducts research and as one examines research conducted by others. Each of the pathways between the boxes in Figure 1, and each of the boxes, represents a series of decisions made by the researcher. Thus, for example, the pathway from Box A to Box B indicated by Arrow 1 (“aspects of the situation are selected as the constructs of importance in the conceptual model”) offers a reminder that any choice of focal phenomena represents a set of theoretical commitments. This provides the opportunity to reflect on the choice and implications of the conceptual model that is being (even if tacitly) employed. For example, which phenomena are not taken into account by this perspective? Which are given significant emphasis? How are those theoretical biases likely to shape the interpretation of the situation?

Similarly, the pathway between Boxes B and C indicated by Arrow 2 (“aspects of the conceptual model are captured in a representational system”) represents an act of *data selection and reduction* as well as representation. In historical studies, for example, whose voices are selected and heard? Or, suppose one is conducting classroom observations. Does one take field notes or make videotapes? If one tapes, what is the focus of the camera? If one takes notes, are they structured according to a predetermined system (in which case they reflect an explicit focus on particular aspects of the situation) or are they somewhat open (in which case the selection is tacit)? For example, data-gathering during the days of the process-product paradigm typically consisted of tallying certain kinds of behavior (teacher actions, student actions) and looking for correlations with educational outcomes (e.g., test scores). In contrast, many current studies of classroom discourse focus on the character of student and teacher interactions, and the

---

<sup>6</sup> If space permitted I would include a third example. Suppose one wanted to conduct an ethnographic study of classrooms using different curricula, with a focus on (say) discourse structures and their impact. It is left as an exercise for the reader to work through Figure 1, with regard to issues such as unit of analysis, selection and form of data, outcome measures (e.g., test scores, or discussions of identity), and interpretation. All of the issues that arose in the quantitative example arise here as well.

results in terms of community norms, beliefs, and knowledge. Each act of data selection, reduction, and representation will have the potential to illuminate certain aspects of a situation, and to obscure others (or even render them invisible). Even if the selection is done with great fidelity to the theoretical model, an act of sampling is taking place.

The third arrow, “analyses are performed within the representational system,” is deceptively simple. The key questions to ask are, What is meaningful within the representational scheme? What can be said about the quality of the inferences drawn? It should be obvious that great care must be taken in subjective analyses. But it is equally important to take comparable care in the case of ostensibly objective quantitative analyses. The results of data analyses will be no better than the quality of the data that are subjected to analysis. For example, there may be a statistically significant difference in the performance levels of two classes on an outcome measure. But is the cause a difference in the two instructional treatments, the fact that they were taught by different teachers, or (if the same teacher taught both) either the enthusiasm of the teacher for one treatment over the other or the fact that one course was taught in the morning and the other right after lunch? Many of the variables that affect performance go unmeasured in statistical analyses. I shall review the issue of *trustworthiness* of analyses in the next section.

The fourth arrow is the mirror image of the second. Just as the passage from a conceptual model to a representational system involves data selection and reduction, the return from the representational system to the conceptual model involves significant acts of interpretation. A difference in two measures might be statistically significant, for example – but is it meaningful or consequential? If so, along what lines? Or, to take a qualitative example, suppose the representational system involves coding student-to-student dialogue in classroom interactions. If the coding scheme focuses on the frequency of interactions and dialogic “take-up,” one might, for example, get a picture of a highly collaborative working group. But what was the collaboration about? An analysis of the content of the interactions might or might not indicate that the group was focused productively on important mathematical issues. Thus the extrapolation from representational system to the conceptual system must be made with care.

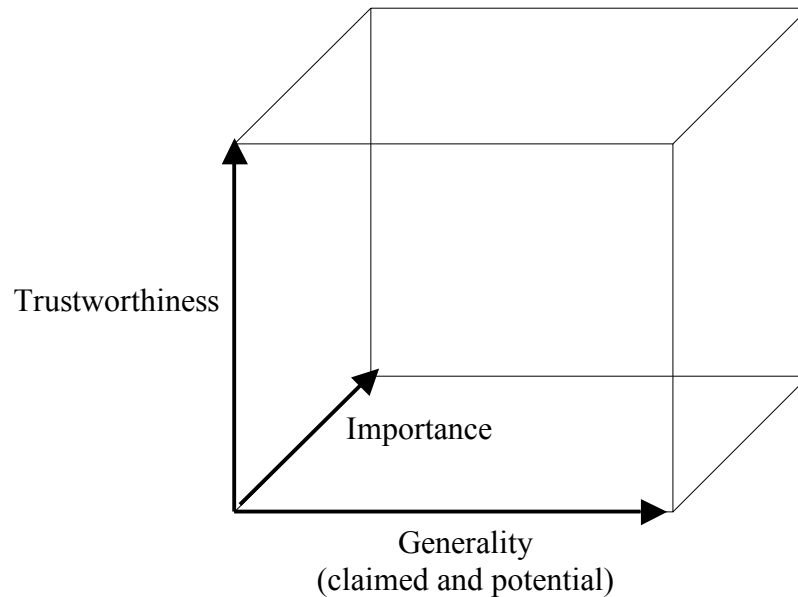
Finally, there is the return from the conceptual model (Box E) to the “real world” situation – the original Box A, now Box F. Here too there is at least interpretation, and perhaps extrapolation. For example, what are test scores taken to mean? History has made clear the consequences of confusing test scores such as IQ tests with the traits they ostensibly represent, such as “intelligence.” Likewise, whether one attributes mathematical proficiency to a good score on the SAT-9 or the Balanced Assessment tests can make a big difference. And, saying “students from Curriculum X did (or did not) outperform students from Curriculum Y on this test” is a very different thing than saying “Curriculum A is (or is not) better than Curriculum B.” I address the idea of *generality* in the next section.

## **Part 2: Aspects of research – issues of trustworthiness, generality, and importance.**

In this section I discuss three fundamental issues related to all research studies. Those issues can be posed as questions that can be asked about any study:

- Why should one believe what the author says? (the issue of trustworthiness)
- What situations or contexts does the research really apply to? (the issue of generality, or scope)
- Why should one care? (the issue of importance)

The following diagram may be useful in thinking about the ultimate contributions made by various studies or bodies of studies:



*Figure 3.* Three important dimensions along which studies can be characterized. (Reprinted, with permission, from Schoenfeld, 2002, p. 450)

As an example, a superbly written and insightful autobiographical account might score high on the trustworthiness and importance dimensions, while low on generality – although it might, by virtue of elaborating what might be seen as common experience, have intimations of generality in some ways. The same could be said of some rich descriptions of rather singular events, such as Fawcett’s (1938) description of a 2-year-long geometry course. Fawcett’s course served as an important and well-documented existence proof: It *is* possible to teach a course in which students develop certain kinds of understandings of the mathematical enterprise. A technically proficient comparison of two instructional treatments might rank reasonably well on the trustworthiness dimension. Such a study might or might not be important, and might or might not have potential generality, depending on the particulars of the situation. Myriad dissertation studies with conclusions of the form “students using materials that I developed scored better than students using a standard text” score low on both. However, individual and collective evaluations of some of the newer and widely used mathematics curricula begin to suggest generality and, as the findings mount, some importance (see, e.g., Senk & Thompson, 2003). Studies that are technically proficient but flawed along the vertical pathways illustrated in Figure 1 are *not* trustworthy. By considering the three dimensions

in Figure 3 one can get a sense of what studies can make as their contributions. In what follows I address the three dimensions one at a time.<sup>7</sup>

### *Trustworthiness*

“What did the President know, and when did he know it?”  
United States Senator Sam Ervin, during the Watergate  
hearings.

The United States Senate’s impeachment hearings of President Richard Nixon, known as “the Watergate hearings,” were one of the defining moments of the American presidency. Richard Nixon’s presidency was at stake. Time and time again, Senator Sam Ervin boiled things down to their evidentiary core. The answer to his oft-repeated question “What did the President know, and when did he know it?” would, once arrived at with a significant degree of certainty, determine Nixon’s fate as president.

What mattered in Nixon’s impeachment trial – what matters in all legal proceedings – is the idea of “a significant degree of certainty.” Legal matters, unlike matters of mathematics, are not axiomatic. Typically, one does not resolve complex legal issues with absolute certainty; rather, the standard is whether particular claims have been substantiated “beyond a reasonable doubt.” The underlying evidentiary issues are: What warrants are given for the claims being made? How believable and trustworthy are they? How robust are the conclusions being drawn from them?

Thus it is in mathematics education as well. As indicated in the previous section, once one has developed a theoretical orientation toward a situation, the core processes of empirical research are the gathering, representation, manipulation, and interpretation of data. A core question regarding the quality of the conclusions drawn from the research is, how *trustworthy* are each of those processes?

In what follows I elaborate on a number of criteria that are useful for examining the quality of empirical and theoretical research. Aspects of trustworthiness include the following, which are discussed below:

- Descriptive and explanatory power
- Prediction and falsification
- Rigor and specificity
- Replicability
- Triangulation

*Descriptive and explanatory power.* *Descriptive power* denotes the capacity of theories or models to represent “what counts” in ways that seem faithful to the phenomena being described. Descriptions need not be veridical but they must focus on what is consequential for the analysis. To give a classical mathematical example, consider a typical related-rates problem that involves a ladder sliding down the side of a building. The building is assumed to be (or is explicitly stated to be) vertical and the

---

<sup>7</sup> The three aspects of research I have called “dimensions” are not truly independent, of course. What I am offering here is a heuristic frame and an argument that attention to all three aspects of research is essential.

ground horizontal. In the diagram representing the situation, the ladder, the building, and the ground are represented as *lines* that comprise parts of a right triangle. What matters for purposes of the desired analysis are their lengths, and the way the ladder is moving. That information, properly represented and analyzed, enables one to solve the given problem; that information and nothing else is represented in the diagram and the equations one derives from it. What does not matter (indeed, what would be distracting in this context) includes how many rungs are on the ladder or how much weight it might support. In a different context, of course, such things would matter quite a bit. The issue of descriptive power, then, is, does the research focus on what is essential for the analysis, in a way that is clear and compelling?<sup>8</sup>

*Explanatory power* denotes the degree to which a characterization of some phenomenon explains how and why the phenomenon functions the way it does. Before getting down to educational specifics, I note that explanatory power, like many of the properties described in this section of this chapter, is an issue in all analytic fields. In mathematics, many people will prefer a constructive proof to an existence argument that employs proof by contradiction. The latter says that something exists, but not how to build or find it; the former, in providing a “blueprint” for finding or building it, provides more detailed guidance regarding how and why it exists.

A fundamental issue here is the distinction between correlation – the fact that X and Y tend to co-occur – and what I shall call explanation at a level of mechanism, an attempt to say how and why X and Y are linked. (Note that causal or constraint-based explanations are the traditional forms of explanation at a level of mechanism, but that statistical or probabilistic models also provide such explanations.)

As one example, consider Mendelian genetics as an expression of a theory of heritability. Surely, before Mendel, the notion that traits appear to be passed from generation to generation had been observed. Indeed, the observation had some power – but that power was limited. With Mendel came a suggestion of mechanism – the idea that (at least in the case of specific traits of relatively simple plants such as peas) genes determine the characteristics of those traits. And with that suggestion came the possibility of experimentation and the refinement of the underlying theoretical notions.

Productive, data-based explanations need not be right in all detail – but they do need to be taken seriously in order for the field to progress. For example, Galileo (at great personal cost) advanced a solar-centric theory of planetary motion, which was more accountable to data and explanation than the faith-based earth-centric model it supplanted. Centuries later this work was established on a much more solid footing when Newton proposed a theory of gravitation that had the potential to take things further. His theory explained, for example, why planetary orbits are elliptical rather than circular. (To anticipate arguments made later in this section, it should also be noted that issues of prediction and falsifiability make a big difference. They allow for the evaluation of

---

<sup>8</sup> As I stress in Part 1 of this chapter, a great deal of what one attends to in empirical research depends on the conceptual-analytic frameworks that orient one to the phenomena at hand. Thus, when I say that descriptive power denotes the capacity of theories or models to represent what counts, I am making a contextual statement: a characterization or representation that has substantial descriptive power has that power relative to the conceptual-analytic framework being employed. Finding a “better description” may entail finding an alternative conceptualization.

competing theories, and the refinement of explanations. And, to emphasize a point made earlier, increased explanatory power may come from a shift in conceptual-analytic frameworks.)

One example within mathematics education concerns the relationship between teachers' classroom practices and student learning. In the 1970s and 1980s a dominant approach in classroom research was the process-product paradigm, in which the data gathered focused on (a) tallies of specific classroom practices (e.g., time on task, worksheet use, asking certain kinds of questions); (b) student outcomes, typically as measured on standardized tests; and (c) the statistical relationships between (a) and (b). The results of such analyses were typically of the form "students did better when teachers did X more frequently," with the implication that it would be good for teachers to do X more often. Of course, researchers had ideas as to why some practices were more productive than others – but the research methods themselves did not explore how or why they worked.<sup>9</sup>

In contrast, more recent studies of classroom activities focus on notions such as discourse communities, practices, and classroom norms and their impact – students' participation structures, their sense of the mathematics, and their understanding as reflected on a variety of assessments. Boaler (2002), for example, described in detail the philosophy and teaching practices at two schools, and the beliefs and understandings they are likely to engender. She interviewed students about their understandings regarding the nature of mathematics and their sense of the in-school enterprise. She examined their performance on standardized tests and on problem-solving tests of her own construction. As a result, readers understand that there are differences – and they have been provided evidence that substantiates a plausible story about how and why those differences came to be.

I should stress that I do not wish to privilege any form of explanation in this discussion. In a field that draws from the social sciences, the sciences, and the humanities, there will be various ways to try to explain how and why things happen. What is important is that the attempt be made and that claims must be held accountable to data. Thus, if one claims that teachers' decision making is based on their knowledge, goals, and beliefs (e.g., Schoenfeld, 1998), one should offer models of that decision making and enough evidence to warrant such claims; if one claims that a particular kind of action or sequence of actions on the part of a teacher supports the development of an

---

<sup>9</sup> Indeed, subsequent studies showed some of the limitations of this kind of approach. In one study comparing highly effective teachers with other teachers, Leinhardt (1990) showed that the highly effective teachers (defined as those teachers whose students performed in the top performance range for their demographic group – whether "learning disabled," "gifted," or anything in between) uniformly established clear disciplinary routines for their classes at the beginning of the school year. In a companion study, however, Leinhardt et al. (1991) documented the limited utility of such findings. It turned out that many of the teachers had rather shaky mathematical knowledge, extending only a little bit beyond that of the curriculum; and that the standardized tests used to assess students (and teachers) were rather narrow and procedural. Thus, the other work could be reinterpreted as indicating that a high level of discipline is effective in helping teachers remain within their "comfort zone" and tends to produce students who have mastered procedures but may have little or no conceptual understanding.

“oppositional culture” in the classroom, one should offer evidence of the growth of opposition and link that growth in plausible ways to the teacher’s actions.

*Prediction and Falsification.* I first describe issues of prediction and falsification in general, then with regard to educational research.

In the physical and other sciences, prediction has been the name of the game, and the potential for falsification a theoretical essential. Physics is the archetype, of course (and, alas, the source of much inappropriate positivism in education). Newton’s laws, for example, say that under certain conditions, certain things will take place. Those predictions, and myriad others, serve both practice and theory. In practical terms, they allow people to build things that work consistently and reliably. Predictions are a mechanism for theoretical progress. As noted above, for example, a solar-centric explanation of planetary motion ultimately triumphed over the earth-centric view because the former explanation fit the data better than the latter. A theory of gravitation based on Newtonian mechanics and the inverse square law of gravitational attraction provides close descriptions and predictions of planetary motion, and allows for fine-grained predictions of planetary movement. The Newtonian view prevailed for centuries, but with some known anomalies – some of its predictions were not quite right. In one of the more famous incidents in the history of science, Einstein’s theory of relativity predicted that under the conditions of a solar eclipse, the planet Mercury would appear to be in a different location than Newtonian theory predicted. It took some years before the eclipse took place – and Einstein’s view was substantiated.

The point, as noted by Popper (1963), is that Einstein could have been proved wrong. Had Mercury not been where Einstein predicted, there would have been evidence of problems with Einstein’s approach. That is, Einstein’s theory was *falsifiable*. Evidence could be gathered that substantiated it, but also that cast it into doubt. According to Popper:

- Every “good” scientific theory is a prohibition: it forbids certain things to happen.
- A theory which is not refutable by any conceivable event is non-scientific.
- Every genuine test of a theory is an attempt to falsify it, or to refute it. Testability is falsifiability; ...
- Confirming evidence should not count except when it is the result of a genuine test of the theory; and this means that it can be presented as a serious but unsuccessful attempt to falsify the theory;
- One can sum up all this by saying that the criterion of the scientific status of a theory is its falsifiability, or refutability, or testability. (Popper, 1963, p. 36)

The kind of predictions made in classical physics represent only one type of prediction, which has given rise to many misconceptions about predictions in the social sciences. Although “absolutist” arguments in domains such as physics may have face validity, theorists such as Toulmin (1958) argue that they do not apply in more complex, contextual situations in which human actions are involved. Toulmin seeks ways to mediate between absolutism on the one hand and relativism on the other; theorists such as

Pickering (1995) replace notions of absolutism with concepts of scientific *practices* that are bound to the histories of the scientific communities in which they emerge. Simply put, the notions of prediction and falsification are unsettled. For that reason, I outline a range of prediction in the sciences and then education, while trying to preserve what I can of the notion of theory-testing.

As noted, the physical sciences sometimes support predictions of the type “under these conditions, the following will take place.” All of the traditional laws in the sciences afford such predictions. For example, the ideal gas law  $PV = nRT$  says that under certain conditions, the values of three of the variables  $P$ ,  $V$ ,  $n$ , and  $T$  determine the value of the fourth. Likewise, the creation of the periodic table as a theoretical characterization of atomic structure supported predictions about the existence of elements that had not yet been found. These kinds of predictions are deterministic.

The life sciences often entail predictions that are not deterministic in the sense above, but which still describe an expected state. Consider, for example, predator-prey models of animal populations. In simplest terms, predators flourish when they are few in number and there are many prey, but they diminish in number when they are densely crowded and few prey remain as food. Prey flourish when they are few in number and predators are few in number, but they diminish in number when they are overcrowded or the number of predators is large. All of these states, and population change rates, can be quantified, at which point the predator-prey model will predict changes in the sizes of both populations. The fates of individual animals are not determined in such models, but trends are. In this sense, predictions are not absolute: A theory is not “true” or “false” in the sense that it proposes a universal, and one counterexample serves to invalidate it. Nonetheless, the theory does give rise to models, and the accuracy of the models (and the theory that generated them) can be judged by their fidelity to actual data. Similarly, Mendelian genetics predicts the percentages of offspring that will have specific traits, but not (except in the case when  $p = 1$ ) the traits of individuals; it yields probability distributions regarding the traits of individuals. Yet, at least cumulatively, this is a strong form of prediction. (And the predictions led to refinements of the theory – for example, data that did not conform with theory led to the uncovering of linked traits.)

A weaker form of prediction has to do with *constraints*. Here, evolutionary theory is a primary example. Under ordinary circumstances, evolutionary theory cannot specify how an organism will evolve – just that it will be responsive to its environment.<sup>10</sup> However, evolutionary theory does impose constraints about the ways on which organisms change over time: There is, for example, greater complexity and differentiation. Thus, according to the theory, certain evolutionary sequences are plausible and others are implausible (if not impossible). Thus the theory can be challenged by empirical evidence. Every exploration of geological strata offers potential disconfirmation.

---

<sup>10</sup> There are exceptions in simple cases of natural selection, where the relationship between certain animal traits and the environment is clear. In one classic case, for example, it was possible to predict that as pollution darkened local trees, the population of moths would darken because darker moths on trees were less visible to predators than lighter moths. (And, when pollution was reversed, the population balance changed back in the other direction.)

And then, of course, there is weather prediction, which is not yet a science. However, various models of climatic behavior can be assessed and refined. The main point here is not that those in the business of building climatic models “have it right” – far from it. But, because each model allows for predictions, and variations in the models support different predictions, the careful examination of the predictions made and their relation to the theory can help to improve prediction and the underlying theory.

I now turn to educational issues. Of course, only some empirical work in the social sciences or in education involves prediction; substantial bodies of empirical work in education (e.g., autobiographical reports and descriptive studies) involve no claims beyond those made about the evidence discussed. However, I note that a great deal of descriptive work contains implicit claims of generality, and thus of prediction. As soon as there is the implication that “in similar circumstances, similar things happen,” one is, at least tacitly, making a prediction (see the section on *generality* below).

Within educational research, as in the examples from the sciences discussed above, there is a wide range of prediction. Randomized controlled trials offer one kind of prediction: The assumption underlying experimentation is that under conditions similar to the circumstances of experimentation, results similar to the results of experimentation will be obtained. This is not unique to statistically oriented studies, however: The same is often true of “rich, thick” anthropological descriptions. A main purpose of descriptions of productive classroom discourse structures is to explain not only how things took place (descriptive power) but why students learned what they did (explanatory power), thus enabling others to try similar things in the hope of obtaining similar results (prediction). The more that such descriptions and claims can be made rigorous, the more likely they are to have a productive impact on practice and to serve theory refinement.

Here are some illustrative examples. Brown and Burton’s (1978) study of children’s arithmetic “bugs” described the authors’ analyses of children’s errors in base-ten subtraction. Brown and Burton found that children’s patterns of errors were so systematic (and rule-based) that, after giving children a relatively short diagnostic test, they could predict with some regularity the incorrect answers that those students would produce on new problems. Brown, Burton, and colleagues (Brown & Burton, 1978; Brown & VanLehn, 1982; VanLehn, Brown, & Greeno, 1984) provided well-grounded explanations of why students made the mistakes they did. But prediction added a great deal to their work. First, the data provided clear evidence of the power of cognitive models: If you can predict the incorrect answers a student will produce on a wide range of problems before the student works them, you must have a good idea of what is going on inside the student’s head! Second, prediction played a role in theory refinement: If predictions do not work, then one has reason to look for alternative explanations. Third, a fact often overlooked by those who view the “buggy” work as overly simplistic and mechanistic is that this work provided clear empirical evidence of the constructivist perspective. In 1978, many people believed in the simplistic idea that one teaches something (perhaps in multiple ways) until the student “gets it,” and that nothing has been “gotten” until the student has learned whatever it was to be learned. Brown and Burton showed that students had indeed “gotten” something: They had developed/learned an incorrect interpretation of what they had been taught and used it with consistency. That is, what they did was a function of what they perceived, not simply what they had

been shown. If that isn't data in favor of the constructivist perspective, I don't know what is. (See also Smith, diSessa, & Roschelle, 1993/1994.)

Another famous study, predating Brown and Burton, is George Miller's 1956 article "The Magical Number Seven, Plus or Minus Two." After numerous observations in different intellectual domains, Miller hypothesized that humans have the following kind of short-term memory limitation: We can only keep between (roughly) five and nine things in working memory at the same time. That hypothesis gives rise to simple and replicable predictions. For example, carrying out the multiplication

$$\begin{array}{r} 634 \\ \times 857 \\ \hline \end{array}$$

requires keeping track of far more than nine pieces of information. According to Miller, it would be nearly impossible to look at these numbers, close your eyes, and compute their product. This has been shown time and time again – once at a talk I gave to more than a thousand mathematicians, many of whom, at least, can do their arithmetic. (It *is* possible to do such products if one knows certain arithmetic shortcuts, or if one rehearses some of the subtotals so that they become "chunked" and only use up one "slot" in short-term memory. But such instances are rare.) Note that the element of falsifiability of Miller's work is critically important. If a nontrivial fraction of the people given tasks like this succeeded at them, then Miller's claim would have to be rejected or modified.

A third class of examples consists of process models of cognitive phenomena. In physics, diSessa's models of *phenomenological primitives* and the ways they develop and shape cognition provide predictions of the ways people will interpret physical phenomena, and tests of the adequacy of the theory. In a different domain, the Teacher Model Group at Berkeley (see, e.g., Schoenfeld, 1998) constructed detailed models of a range of teachers' decision making. One theoretical claim made by the group is that a teacher's in-the-moment actions can be modeled as a function of the teacher's knowledge, goals, beliefs, and a straightforward decision procedure. Each new case explored (an inexperienced high school mathematics teacher conducting a more or less traditional lesson, an experienced high school teacher conducting an innovative lesson of his own design that largely went "according to plan," and an experienced third-grade teacher conducting a lesson in which the agenda was coconstructed with the students) tested the scope and adequacy of the theory. Again, the idea is straightforward. The more that theoretical claims can be examined and tested by data, the more there is the potential for refinement and validation.

*Rigor and specificity.* At this point, little needs to be said about the need for rigor and specificity in conducting empirical research. The more careful one is, the better one's work will be. And the more carefully one describes both theoretical notions and empirical actions (including methods and data), the more likely one's readers will be able to understand and use them productively, in ways consistent with one's intentions.

Precision is essential; a lack of specificity causes problems. One historical example is the potentially useful construct of *advance organizer* introduced by Ausubel (see, e.g., Ausubel, 1960). Over the years, a large number of studies indicated that the use of advance organizers (roughly speaking, top-level introductions to the content of a body

of text) improved reading comprehension. But an equally large number of studies indicated that the use of advance organizers did not make a difference. Why? On closer examination, the construct was so loosely defined that the advance organizers used in various studies varied substantially in their characteristics. Thus the results did as well. Similarly, a theory that involves terms such as *action, process, object, and schema* needs to say what those terms are, in clear ways. And, of course, appropriate detail and warrants need to be provided in discussing those terms. One example of a productively used term is the *didactical contract* as used by the French (see, e.g., Brousseau, 1997). The term has a specific meaning, which provides a useful backdrop for many such studies. In contrast, the generic use of the term *constructivist* (for example, in the meaningless term *constructivist teaching*) has not been helpful.

Note that rigor does not simply mean rigor in the use of one's data analyses (the path denoted by Arrow 3 in Figure 1). It means attending carefully to all of the pathways from Box A to Box F in Figure 1.

*Replicability.* Every person is different; every classroom is different. How can one possibly speak of replication in education? The idea seems strange. One might replicate experiments of some types – but how many educational researchers do experimental work? And, if every classroom is different, what does it mean to replicate someone else's research study?

One way to think about the issue of replicability is to think about generality. As noted above and as will be elaborated below, there is a tacit if not explicit aspect of generality to most studies: the expectation is that the lessons learned from a study will apply, in some way, to other situations. To that degree, some key aspects of those studies are assumed to be replicable. The issue, then, is, how does one characterize those aspects of a study – in enough detail so that readers can profit from the work in the right ways, and so that they can refine the ideas in it as well? Thus, for example, studies like those referred to in the section on prediction (Brown and Burton's studies of arithmetic bugs, Miller's article on the magic number seven plus or minus two, and the Teacher Model Group's article on teachers' decision making) all involve prediction; hence they should be potentially replicable as well.

*Multiple sources of evidence (triangulation).* In a well-known article published in 1962, Martin Orne introduced the concept of the *demand characteristics* of an environment via this personal anecdote:

A number of casual acquaintances were asked whether they would do the experimenter a favor; on their acquiescence, they were asked to perform five push-ups. Their response tended to be amazement, incredulity and the question 'Why?' Another similar group of individuals were asked whether they would take part in an experiment of brief duration. When they agreed to do so, they too were asked to perform five push-ups. Their typical response was "Where?"

The idea, later abstracted as *context effects*, is that context makes a difference: People will do things in some circumstances that they might do differently (or not at all) in other circumstances. Some behavior is purely artifactual, *caused by* the particulars of the circumstances.

For this reason, the use of multiple lenses on the same phenomena is essential. In some cases, that means employing multiple methods to look at the same phenomena. Thus, observations, questionnaires, and interviews can all be used to challenge, confirm, or expand the information gathered from each other. Similarly, behavior that manifests itself in some contexts may or may not manifest itself in other contexts.<sup>11</sup> One should constantly be on guard with regard to such issues.

### *Generality and Importance*

A central issue with regard to any research is its scope, or generality – the question being, how widely does this idea, or theory, or finding, actually apply? A second, equally critical issue is importance. A study may or may not be warranted to apply widely. But whether it is or is not, one must ask: Does it matter? Just what is its contribution to theory and practice?

The concept of generality is slippery, because in many papers a few instances of a phenomenon are taken as an indication of the more widespread existence (or potential existence) of that phenomenon. Thus, it may be useful to distinguish the following kinds of generality:

- The *claimed generality* of a body of research is the set of circumstances in which the author of that work claims that the findings of the research apply.
- The *implied generality* of the work is the set of circumstances in which the authors of that work appear to suggest that the findings of the research apply.
- The *potential generality* of the work is the set of circumstances in which the results of the research might reasonably be expected to apply.
- The *warranted generality* of the work is the set of circumstances for which the authors have provided trustworthy evidence that the findings do apply.

There is often a significant gulf between these. And there is often slippage between evidence-based claims (warranted using the constructs within the conceptual-analytic system – the path from Box C to Box D in Figure 1) and explicit or implicit claims of generality (the path from Box A to Box F in Figure 1). Thus, for example, an experimental study may compare outcomes of a new instructional method with an unspecified control method. If the study is done by the author using a measure that he or she developed for the purpose of this study, the warranted conclusion may be “students in this context, taught by the instructor who developed the materials and tested using an assessment similar in kind to the materials, did better on that assessment than students who took a somewhat different course.” The warranted generality is thus quite low. The claimed generality might be “students in the experimental group outperformed control students,” with the implied generality being that other students who experience the new instructional method will similarly outperform other students. The potential generality is actually unknown – it depends on the contexts, the character of instruction in experimental and control groups, the attributes of the assessment, and more.

---

<sup>11</sup> *Context* is meant to be taken very broadly here. Orne’s example is a case in point: a question asked of a friend is very different from a question asked of a voluntary participant in an experiment. Similarly, a question asked of individual experimental participants may be treated very differently by *pairs* of subjects.

A key point is that the warranted generality of a study and its importance are not necessarily linked. For example, Wilbur and Orville Wright demonstrated (with substantial trustworthiness!) on December 17, 1903, that a heavier-than-air machine could take flight and sustain that flight for 59 seconds. That was the warranted generality. However, that flight was critically important as an existence proof. It demonstrated that something *could* be done and opened up fertile new territory as a result. Likewise, many studies in mathematics education can be seen as existence proofs. They may demonstrate the existence of a phenomenon worthy of investigation (e.g., early studies of metacognition or beliefs) or of instructional possibilities (e.g., early studies of Cognitively Guided Instruction). The findings of studies with existence proofs are not yet general – but there may be the potential for them to be.

Conversely, a body of research that has broad generality can turn out to be theoretically sterile and to have little practical value. One such example is a spate of studies conducted in the 1980s, which showed that a substantial proportion of those examined produced the equation  $P = 6S$  instead of  $S = 6P$  in the 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.”

(Clement, 1982; Clement, Lochhead, & Monk, 1981; Rosnick & Clement, 1980). The phenomenon was documented repeatedly, and various attempts were made to get people to do better at such problems. However, after all was said and done, the field had no real understanding of why people made this mistake, and no effective mechanisms for preventing or fixing it. Thus – given that this body of work was general, but did not produce significant understandings or applications – we see that generality and importance are at least somewhat independent dimensions.

In what follows I briefly describe the (warranted-to-potential) generality of a number of studies and discuss the importance of those studies. Broadly speaking, the studies are clustered by their degree of generality. For the most part I have chosen studies whose trustworthiness is well established. Judgments of importance reflect the author’s perspective but are also included as a reminder that importance is an essential dimension to take into account.

#### *Studies of limited warranted generality*

There is a large class of studies for which there is limited warranted generality, but which are worth noting because of their current or potential importance. Such studies may be of interest because they offer existence proofs, bring important issues to the attention of the field, make theoretical contributions, or have the potential to catalyze productive new lines of inquiry.

An autobiographical report may be of interest because it is motivational, or, like other historical documents, it illuminates certain historical decisions or contexts. (For example, what was the historical context for the *Brown Versus Board of Education*

decision?<sup>12</sup>) In cases such as these there are no warranted conclusions beyond those described, but there may be lessons to be learned nonetheless. (Note that there is an implied aspect of generality to historical studies in Santayana's oft-quoted statement that "those who cannot learn from history are doomed to repeat it.")

Another singular event is the existence proof – a study that shows that something is possible and may elaborate on the means by which it became possible. For example, Harold Fawcett's classic 1938 volume *The Nature of Proof* demonstrated that it was possible to create a classroom mathematical community in which students engage meaningfully in many of the activities in which mathematicians engage (e.g., making definitions and deriving results logically from them). Early papers on Cognitively Guided Instruction (CGI) showed that it is possible to capitalize on young students' informal models and situational understandings to help them produce a wide range of meaningful solutions to word problems (see, e.g., Carpenter, Fennema, & Franke, 1996). Moll, Amanti, Neff, and Gonzalez (1992) showed how it is possible to develop instruction that builds on the kinds of cultural knowledge and traditions that students have in their out-of-school lives. Gutstein (2003) showed that it is possible to teach a mathematics course in which students do well on traditional measures of mathematical performance and become engaged as social activists as well. Boaler (in press) showed that it is possible to create a discourse community in a high school mathematics classroom with a large percentage of low-SES and ESL students in which there are high mathematical standards – and the classroom accountability structures are such that the students hold each other accountable for producing meaningful and coherent mathematical explanations. Each of these studies had a high level of trustworthiness, in that it satisfied many of the criteria discussed in the previous section of this chapter. Each had relatively low warranted generality, in that it made the case that something had happened in specific circumstances, with a small number of students. But each showed that something could be done, opening up a previously undocumented space of possibilities. In that sense, they rate high on the potential importance scale (and may, indeed, pave the way to phenomena that become more general). Over time, some (e.g., Fawcett's study) have languished as inspirational examples not taken up on a large scale; some (e.g., Cognitively Guided Instruction) have had increasing impact.

Other studies bring important phenomena to readers' attention. They may suggest as-yet-undocumented generality, opening up arenas for investigation. Thus, for example, Cooney's (1985) case study of "Fred" showed that a teacher may espouse a set of values that sounds compatible with a particular pedagogical direction (in this case, "problem solving") but may interpret those terms in contradictory ways and thus act in ways contrary to (the normative interpretation of) those values. Cohen (1990) showed that a teacher may adopt the rhetoric of a particular pedagogical approach and believe that she is implementing that approach, while in fact assimilating (in the Piagetian sense) many of the surface aspects of that approach to her long-established pedagogical practices. Eisenhart et al. (1993) showed that a particular teacher, subject to the pressures of her environment, wound up teaching in a way differently than she intended. Each of these studies made a well-documented case that something important was happening in one

---

<sup>12</sup> In a 1954 decision that had far-reaching implications, the United States Supreme Court declared that the establishment of segregated public schools was unconstitutional.

specific set of circumstances. But each study also made a plausibility case that the phenomenon under discussion was more widespread.

A classic example of this genre is Heinrich Bauersfeld's 1980 article "Hidden Dimensions in the So-Called Reality of a Mathematics Classroom." Bauersfeld revisited the data from a dissertation by Shirk (1972), which had focused largely on the content and pedagogical goals of beginning teachers. Examining the same data, Bauersfeld focused on the social dimensions of the classroom. His analysis addressed four areas of classroom research that had hitherto received a negligible amount of attention: "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). Although the warranted generality of the findings was small, the face-value typicality of the classroom he explored made at least a plausibility case that the phenomena under investigation were relatively widespread. The phenomena were not (yet) claimed to be general but were seen as worthy of investigation.

In a similar way my 1988 article "When Good Teaching Leads to Bad Results: The Disasters of Well Taught Mathematics Classes" presented a trustworthy account of one classroom, in which students had developed a series of counterproductive beliefs about the nature of mathematics as a result of receiving well-intentioned but narrow instruction that focused on preparing the students with particular procedural skills to do well on a high-stakes test. By virtue of the typicality of the instruction in the focal class, and the causal explanation of how the classroom practices resulted in specific outcomes, the article had substantial potential generality.

Other papers may be important because of their theoretical or methodological contributions. For example, diSessa's 1983 chapter on *phenomenological primitives* introduced a new way to conceptualize the development of conceptual structures and reframed the theoretical debate on the nature of conceptual understanding. Brown and colleagues (e.g., Brown, 1992; Brown & Campione, 1996) introduced and elaborated on the notion of design experiments, describing nonstandard instructional practices and a novel way to think about data gathering and analysis in the context of such instruction. Yackel and Cobb's (1996) discussion of sociomathematical norms provided a theoretical tool (which will have widespread application) for examining the ways in which "taken-as-shared" classroom practices can shape individual cognition. Similarly, Cobb and Hodge's (2002) characterization of diversity as a function of differences in cultural practices rather than a measure of "spread" of some particular demographic variable, although not yet widely used, has the potential to reframe studies of diversity in productive ways. Therein lies its importance.

*Studies where some degree of generality is claimed and/or warranted.* The simplest claims of generality are statistical. Consider, for example, the following datum from Artigue (1999):

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 and of itself, this statement (whose generality is warranted in statistical terms) may or may not seem important. However, this and other data (e.g., that a significant

number of such students believe that the decimal number 0.9999... is infinitely close to but less than 1) help to establish realistic expectations regarding the knowledge of entering calculus students, and expectations about necessary focal points of instruction.

In a similar way, some findings from the U.S. National Assessment of Educational Progress (NAEP) pointed to serious across-the-board issues in American mathematics instruction. Perhaps the best-known single item from NAEP is the following (Carpenter, Lindquist, Matthews, & Silver, 1983):

An army bus holds 36 soldiers. If 1128 soldiers are being bussed to their training site, how many buses are needed?

The calculation is straightforward: 36 goes into 1128 thirty-one times, with a remainder of 12. Hence 32 buses are needed. Here is the frequency of responses from the stratified nationwide sample of students who worked the problem:

23%	32
29%	31R12
18%	31
30%	other

The vast majority of students who worked the problem performed the necessary computation correctly. But then, more than a third of the students who did the computation correctly wrote down an impossible answer (the number of buses needed must be an integer, after all!), and a substantial number of others rounded down instead of up, which makes no sense in the given context.

Once again, the statistics attest to the generality of the finding. What made this finding *important* was the way it fit into an emerging body of research findings. In the early 1980s, researchers were coming to understand the importance of beliefs (Schoenfeld, 1983), both in terms of their impact on students' mathematical performance and in terms of their origins in classroom practices. Many of the arguments in support of beliefs had been local: In a specific class, students experienced these specific things; they developed the certain understandings of the mathematical enterprise as a result; and they then acted in accord with those understandings. In small-scale studies the argument had been made that students had come to experience word problems not as meaningful descriptions of mathematical situations, but as "cover stories" for arithmetic operations; they understood their task as students as uncovering the relevant mathematical numbers and operations in the cover story, performing the operations, and writing the answer down. In this view, the ostensible reality described in the cover story played no role in the problem once the numbers and operations had been extracted from it. The data on the NAEP provided evidence that the problem was a national one. Therein resides the importance of the research.

In a similar way, I first explored the role of metacognition in mathematical problem solving in the research described in part 1 of this chapter. The data in the original experiment were suggestive: more than half of the videotapes I examined were of the type shown in Figure 2. This finding was potentially important. The question was whether it would turn out to be general. The number of cases I had examined was small, but a priori there was nothing to suggest that they might be atypical. When the kinds of

effects demonstrated in my early studies were replicated elsewhere and tied to emerging theories of the importance of metacognition in other domains (e.g., Brown, 1987), the generality and importance of the phenomenon became increasingly clear. Note that this kind of sequence is typical: Individual studies point to aspects of a potentially interesting or general phenomenon, and an expanding body of studies refine the idea over time.

A set of ideas that is in its early stages, suggesting generality but not yet fully explored or validated, has to do with the attributes of particular kinds “learner-centered” learning environments. Engle and Conant have suggested that

productive disciplinary engagement can be fostered by designing learning environments that support (a) problematizing subject matter, (b) giving students authority to address such problems, (c) holding students accountable to others and to shared disciplinary norms, and (d) providing students with relevant resources. (2002, p. 399)

The authors provided a detailed analysis of a series of discussions in a “Fostering a Community of Learners” classroom (Brown & Campione, 1996), demonstrating how each of the four themes identified above played out in that classroom. They then briefly reexamined discourse patterns in two other types of instruction known for engaging students productively with disciplinary content: Hypothesis-Experiment-Instruction Method Classrooms (Hatano & Inagaki, 1991; Inagaki, 1981; Inagaki, Hatano, & Morita, 1998) and the Chèche Konnen project (National Science Foundation, 1997; Rosebery, Warren, & Conant, 1992). This evidence suggests the potential generality of the findings. Readers’ familiarity with other such environments – e.g., Scardamalia and Bereiter’s “Knowledge Forum” classrooms (Scardamalia, 2002; Scardamalia & Bereiter, 1991; Scardamalia, Bereiter, & Lamon, 1994) or my problem-solving courses (Schoenfeld, 1985) may add to the sense of generality and importance of the findings. Time will tell.

A final kind of argument with some generality and potential importance, which is a bit further along than the Engle and Conant study, is done by aggregating over a collection of studies, each of which offers consistent data. Such an argument is made in Sharon Senk and Denisse Thompson’s 2003 volume, *Standards-Based School Mathematics Curricula: What are They? What do Students Learn?* Senk and Thompson present the data from studies evaluating a dozen *Standards*-based curricula. A careful reader could find reasons to quibble with a number of the studies. For example, many of the assessments employed in the studies were locally developed<sup>13</sup>. Hence one could argue that some of the advantages demonstrated by the *Standards*-based curricula were due in part to assessments tailored to those curricula. In addition the conditions of implementation for the curricula (e.g., the preparation that teachers received before teaching the curriculum) may have been higher than one might expect in “ordinary” curriculum adoptions. Nonetheless, the pattern of results was compelling.

The book is divided into sections assessing curricula at the elementary, middle, and high school levels. Putnam (2003) provided the following summary of the

---

<sup>13</sup> A concern about locally developed measures is that there is the potential for bias (giving an unfair advantage to the “experimental” curriculum) whenever the developers of that curriculum are closely tied to the developers of an assessment used in the evaluation of that curriculum. Whether advertently or not, items on the assessment could be designed in ways that favor the experimental curriculum.

evaluations of *Math Trailblazers*, *Everyday Mathematics*, *Investigations*, and *Number Power* (all elementary curricula):

The four curricula ... all focus in various ways on helping students develop conceptually powerful and useful knowledge of mathematics while avoiding the learning of computational procedures as rote symbolic manipulations.

The first striking thing to note about [them] is the overall similarity in their findings.

Students in these new curricula generally perform as well as other students on traditional measures of mathematical achievement, including computational skill, and generally do better on formal and informal assessments of conceptual understanding and ability to use mathematics to solve problems. These chapters demonstrate that “reform-based” mathematics curricula can work. (p. 161)

Chappelle (2003) summarized the evaluations of *Connected Mathematics*, *Mathematics in Context*, and *Middle Grades MATH Thematics: The STEM Project* (all middle grades curricula) as follows:

Collectively, the evaluation results provide converging evidence that *Standards*-based curricula may positively affect middle-school students’ mathematical achievement, both in conceptual and procedural understanding.... They reveal that the curricula can indeed push students beyond the “basics” to more in-depth problem-oriented mathematical thinking without jeopardizing their thinking in either area (pp. 290-291)

Swafford (2003) examined the evaluations of the following high school curricula: *Core-Plus Mathematics Project*, *Math Connections*, *the Interactive Mathematics Program (IMP)*, *the SIMMS Integrated Mathematics Project*, and *the UCSMP Secondary School Mathematics Program*. She concluded that:

Taken as a group, these studies offer overwhelming evidence that the reform curricula can have a positive impact on high school mathematics achievement. It is not that students in these curricula learn traditional content better but that they develop other skills and understandings while not falling behind on traditional content. (p.468)

Although many of the individual studies in Senk and Thompson (2003) may be problematic in some regard, the cumulative weight of the evidence in favor of the impact of the *Standards*-based curricula – and the absence from the literature of any comparable evaluations favoring the more traditional, skills-oriented curricula – is such that the findings, although preliminary, acquire a nontrivial degree of trustworthiness. There is a degree of generality in that all of the curricula examined, although differing substantially in style and content, all emphasized conceptual understanding, problem solving, and deeper engagement with mathematical concepts than the “traditional” curricula with which they were compared. And if students learn more, there is a *prima facie* case for importance.

A final example with a somewhat different kind of warrant is the claim made by Stigler and Hiebert that “teaching is a cultural activity (1999, p. 85).” As Stigler and Hiebert noted, the Third International Mathematics and Science Study (TIMSS) video study performed a random sampling of classrooms selected from the larger TIMSS study, which was carefully constructed; the final video study sample included 100 German, 50

Japanese, and 81 U.S. classrooms that the authors claimed “approximated, in their totality, the mathematics instruction to which students in the three countries were exposed” (p. 19). Classroom videos were then coded for various kinds of detail, and the codings were analyzed. The data revealed that there was much less within-country variation than across-country variation – that is, that there were relatively consistent practices in each nation that differed substantially from the practices in the other nations. It goes without saying that one might code for things other than those captured by the coding scheme developed for the TIMSS study. But the main finding, backed up by descriptions of lesson content, appears trustworthy and (by virtue of the sampling) general across the three countries in the study. Subsequent replication studies are extending the findings. But, do the results matter? In this author’s opinion, the answer is a clear *yes*. Cross-national comparisons reveal what turn out to be cultural assumptions that one would not see if one kept one’s eyes within one nation’s borders. They reveal practices that challenge one’s notion of what may be possible and thus challenge one to think about the premises underlying instruction.

*Studies where significant generality, if not universality, is claimed or implied.* On the one hand, as Henry Pollak once said, “there are no theorems in mathematics education.” On the other hand, there are some well-warranted results of significant generality. Many of these have to do with (attributions of) cognitive structures. Thus, for example, although one can certainly quarrel with major aspects of Piaget’s work, phenomena such as children’s development of *object permanence* and *conservation of number* are well established and well documented. Similarly, replication and follow-up studies of phenomena such as “the magic number seven plus or minus two” (Miller, 1956) have established the limitations of short-term memory as essentially universal. Cumulative bodies of literature with regard to the role and existence of schemata or their equivalents as forms of knowledge organization have established the utility of those constructs for describing people’s knowledge structures, likewise for studies of beliefs and metacognition as shapers of cognition. Cumulatively, these studies provide documentation of phenomena that are robust (hence the research is trustworthy); that are essentially universal; that provide an understanding of individual cognition; and are thus clearly important.

Robust findings can, of course, be misused in applications to mathematics education. For example, the fact that people who are fluent in secondary school mathematics possess a large number of schemata for solving word problems does not imply that students should be directly taught a large set of those schemata. (See the exchange between Mayer, 1985, and Sowder, 1985, for a discussion of this issue.) Nor should a focus on cognitive structures in the preceding examples be taken in any way as suggesting that cognitive structures provide complete and coherent explanations of thinking, teaching, and learning. As noted in previous sections, many ideas from sociocultural theory have great promise in helping to unravel the complex interactive processes in which all humans engage. The arena is important, and many individual findings are trustworthy, but the field is still in its genesis. One expects to see many examples with a sociocultural emphasis – studies of discourse patterns, identity formation, etc. – in this category of studies when the third edition of this *Handbook* is published.

Finally, I note another category of potential universals: large-scale comparative studies of various types. As examples, cross-national studies such as TIMSS (see, e.g., Mullis et al., 1998; Mullis et al., 2000) and PISA (see, e.g., Lemke et al., 2004), provide trustworthy comparative information at nationwide levels about student performance. However, it should be remembered that any assessment is only as good as the items it uses (cf. Part 1 of this chapter). Moreover, any assessment reflects the mathematical values of its developers. Specifically, PISA and TIMSS emphasize different aspects of mathematical competence (PISA being more “applied”). Nations that do well on one assessment do not necessarily do well on the other. Other comparative studies (e.g., Lee, 2002) reveal trends in similarities and differences in the mathematical performance of various subgroups of larger populations. The same caveats apply.

### *Discussion*

A major challenge in the conduct of educational research is the tension between the desire to make progress<sup>14</sup> and the dangers of positivism and reductivism. The point of Part 1 of this chapter was that reductivism, at least, comes with the territory: One’s explicit or implicit theoretical biases frame what one looks at, how one characterizes it, how one analyzes it, and how one interprets what one has analyzed. It is easy to slip into positivism as well.

It is in that spirit that I refer to the ideas in Part 2. The question addressed here is, what makes for good empirical work? I have argued that research must be examined along three somewhat independent dimensions. The first is trustworthiness – the degree of believability of the claims made in a piece of research. The core issue is: if claims are made, do the warrants for them ring true? As discussed above, there are various criteria for the trustworthiness of empirical research: a study’s descriptive power; its explanatory power; whether the claims made are in some sense falsifiable; whether the study makes predictions and, if so, how well those predictions fare; how rigorous and detailed the work is; whether the work has been described in ways that allow for attempts at replication and, if so, whether the findings are duplicated or extended; and whether the study offers multiple lenses on the same phenomena and multiple lines of evidence that converge on the same conclusions. Of course, not every criterion is relevant for every study, the nature of a study will determine how each of the relevant criteria is applied. But trustworthiness is an essential quality of good research, and attending to the criteria discussed in Part 2 will help one to make informed judgments about it.

Trustworthiness is not enough, however. Simply put, a study may be trustworthy (and thus publishable in the sense that it meets the appropriate methodological criteria) but trivial, along one or both of the two other dimensions: generality (scope) and importance. Assessing the generality of a result is often a delicate matter. Typically authors imply the generality of a phenomenon by tacitly or explicitly suggesting the typicality of the circumstances discussed in the study. Implying generality is one thing, however, and providing solid evidence for it is another. For that reason I have introduced

---

<sup>14</sup> If one defines *progress* in educational research as either clarifying/adding to the field’s knowledge or producing information or materials that help people do something more effectively, then just about all research is aimed at progress in some way. I intend this broad a definition.

the notions of *claimed*, *implied*, *potential*, and *warranted generality* as ways to think about the scope or generality of a study.

Importance is, of course, a value judgment. But it is an essential one, to be made reflectively.

### **Part 3: From ideas to implementation: A reconsideration of the concept of “clinical trials” in educational research and development.**

My purpose in Part 3 of this chapter is to describe a sequence of research and development activities employing both qualitative and quantitative methods that is intended to serve as a mechanism for the improved development and effectiveness of educational interventions. This effort is motivated both by a need, in general, for mechanisms to improve the curriculum design and evaluation process (Burkhardt, 2006; Burkhardt & Schoenfeld, 2003) and the wish to bring together the statistical and mathematics education communities in profitable ways (Scheaffer, 2006). I shall try to draw upon the best of three traditions that are often seen as in conflict with one another: educational research and design as craft, fundamental research in mathematics education, and calls for the scientific validation of curricular effectiveness through experimental means.

This part of this chapter differs substantially from the two that preceded it. Earlier in this chapter, my goal was generality. Part 1 offered a scheme for characterizing and reflecting on all empirical research, and Part 2 offered a framework that can be used to assess the quality of all empirical research studies. In a sense, the issues addressed in Parts 1 and 2 are timeless – questions of how to conduct research of increasingly high quality will always be with us. Here my focus is more narrow. In a chapter that addresses issues similar to those addressed here, Clements (2002) poses this question: “Why does curriculum development in the United States not improve?” (p. 599) One of my goals is to help solve this practical problem. The problem is timely for practical and political reasons, as will be discussed below. It is far from ephemeral, however. The general issue of experimentalism in education, and the scientism that bedevils it, have been problematic since long before I entered the field.

I begin with some caveats, in the hope of avoiding misinterpretations of what follows. I am *not* proposing that curriculum development is the solution to instructional problems in mathematics education; nor am I suggesting that it is the only (or even the primary) way in which mathematics instruction can be improved. A curriculum is a tool, and a tool is only effective in the hands of those who can wield it effectively. The fact that I do not focus here on issues of professional development, or on strategies for improving instruction via professional development, should not be taken as an indication that I underestimate the value of professional development as a vehicle for improving instruction. In theoretical terms, I do not propose that curriculum development is the solution to the problem of “travel” – the ability of materials developed in one context to be used profitably in another (National Academy of Education, 1999). The more narrow issue that I address is, how can we refine the curriculum development process, so that the curricular tools that are developed are more likely to be effective (when used in the right ways)? The approach I take is to try to develop a serious educational analog to the idea of

clinical trials in the development of medicines and medical treatments. The general approach outlined here should be applicable to all educational interventions.

I note by way of preface that I have been engaged in educational research for long enough to see a number of pendulum swings in what appears to be the eternal (and erroneous) dialectic between emphases on qualitative and quantitative methods. In the 1960s and 1970s, an emphasis on experimental and statistical methods was stifling the field. The problem was often a misapplication of good statistical methods. In many cases a researcher employed a “Treatment A *versus* treatment B” experimental design to compare curricular materials developed by the researcher with some form of control treatment. The assessment measures employed were typically standardized tests or measures developed by the author, each of which have their own problems (recall the discussion in Part 1 of this chapter). Sometimes the experimenter taught both the experimental and control classes, in which case the most significant (and unmeasured!) variable may have been teacher enthusiasm for the experimental materials, or the fact that one course was taught before lunch and one after. Sometimes different teachers taught the experimental and control classes, in which case teacher variation (again unmeasured) may have been the most important variable shaping outcomes. A number of scholars, including myself (see also Kilpatrick, 1978), have argued that there was as much scientism as science in educators’ attempts to adopt scientific methods in educational research – for example, I called for a moratorium on the use of factor analyses on “mathematical abilities” until researchers could explain what the factors determined by statistical methods actually meant.

Over the 1980s and 1990s a flowering of the cognitive and sociocultural perspectives led to a wide range of methods and approaches. Attention to these, combined with a reaction against the sterility of the earlier decades of “scientific” work, led to a general abandonment of quantitative research in the field. This too is problematic, unless there is a firm commitment on the part of researchers to provide the warrants that allow research to be deemed truly trustworthy (cf. Part 2 of this chapter). In a column leading up to the American Educational Research Association annual program in 1999, program chair Geoffrey Saxe and I (1998) reproduced the following quote from a letter I had received:

At [Annual Meetings] we had a hard time finding rigorous research that reported actual conclusions. Perhaps we should rename the Association the American Educational Discussion Association. . . . This is a serious problem. We serve a profession that has little regard for knowledge as the basis for decision making. By encouraging anything that passes for inquiry to be a valid way of discovering answers to complex questions, we support a culture of intuition and artistry rather than building reliable research bases and robust theories. Incidentally, theory was even harder to find than good research. (p. 33)

At the federal level, at least, the pendulum has swung back in the direction of experimentalism in recent years. Here is an excerpt from the U.S. Department of Education's (2002) Strategic Plan for 2002-2007:

Strategic Goal 4: Transform education into an evidence-based field

Unlike medicine, agriculture and industrial production, the field of education operates largely on the basis of ideology and professional consensus. As such, it is subject to fads and is incapable of the cumulative progress that follows from the application of the scientific method and from the systematic collection and use of objective information in policy making. We will change education to make it an evidence-based field. We will accomplish this goal by dramatically improving the quality and relevance of research funded or conducted by the Department, by providing policy makers, educators, parents, and other concerned citizens with ready access to syntheses of research and objective information that allow more informed and effective decisions, and by encouraging the use of this knowledge. (U.S. Department of Education, 2002, p. 51)

Grover Whitehurst, Director of the Department of Education's Institute of Education Sciences, has followed through on that agenda. IES has funded the What Works Clearinghouse, some of whose limitations are discussed in Part 1 and in Schoenfeld (2006a, 2006b). Early in his tenure, Whitehurst made his agenda clear in a presentation (2002) to the American Educational Research Association. Referring to Donald Stokes's argument in *Pasteur's Quadrant* (Stokes, 1997; see below), Whitehurst argued strongly in favor of his Institute's mission to support fundamentally applied work, in "Edison's quadrant":

Yes, the world needs basic research in disciplines related to education, such as economics, psychology, and management. But education won't be transformed by applications of research until someone engineers systems and approaches and packages that work in the settings in which they will be deployed. For my example of massed versus distributed practice, we need curricula that administrators will select and that teachers will follow that distributes and sequences content appropriately. Likewise, for other existing knowledge or new breakthroughs, we need effective delivery systems. The model that Edison provides of an invention factory that moves from inspiration through lab research to trials of effectiveness to promotion and finally to distribution and product support is particularly applicable to education.

In summary, the Institute's statutory mission, as well as the conceptual model I've just outlined, points the Institute toward applied research, Edison's quadrant. (Whitehurst, 2002; unpaginated.)

While paying homage to other forms of research, Whitehurst goes on to state his stance in clear terms: "Randomized trials are the only sure method for determining the effectiveness of education programs and practices."<sup>15</sup> (Whitehurst, 2002)

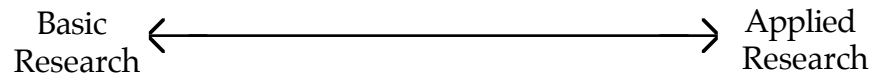
But *sure* in what sense? As noted in the section of Part 1 called "a second example," how one interprets curricular effectiveness depends on how (if at all) one defines and attends to context. And where and when does one move from investigative studies to documentation of effectiveness? Must they be separate? In this section I address these issues.

---

<sup>15</sup> Whitehurst goes on to say that randomized controlled trials (RCTs) are not appropriate for all research questions, and that other methods are appropriate to supplement the findings of RCTs. However, it is clear that "questions of what works are paramount for practitioners; hence randomized trials are of high priority at the Institute."

Because much of the move toward evidence-based research in education is based on comparable work in the health sciences (see, e.g., the work of the Cochrane Collaboration, at <http://www.cochrane.org/>), I shall make my proposal by way of analogy to the model of clinical trials favored in medical and pharmaceutical research. It should be understood, however, that in using this analogy I am not accepting the reductive metaphor of curricular “treatments” in education as being anything like “simple” drug treatments for medical conditions. In what follows I wish to honor the complexity of the human and social processes that characterize the educational enterprise. What I wish to show is that the pharmaceutical enterprise is more complex than many of those who would analogize to education would have it. There is much to learn from the systematic study of proposed medical interventions, *if* the proper analogies are made, fully respecting the complexity of the educational process. If they are not, however – if steps in the process are skipped, or if proper attention is not paid to the complexities of educational research discussed in Part 1 of this chapter – then there is the potential for reductivism or scientism to do more harm than good.

I begin, with some prefatory comments about *Pasteur’s quadrant*. Stokes (1997) observed that, traditionally, basic and applied research have been seen as being at opposite ends of the research spectrum. This is captured by the one-dimensional picture in Figure 4.



*Figure 4.* Basic and applied research viewed as polar opposites

Moreover, basic research has normally been thought of as preceding applied research, which precedes large-scale applications in practice. Stokes argued, however, that the basic/applied dichotomy was too simple. Although some research is almost exclusively basic (the work of physicist Niels Bohr is a classic example) and some is almost exclusively applied (Thomas Edison’s being a case in point), some critically important research has been framed in ways that allowed it to simultaneously make fundamental contributions to theory *and* to the solution of important practical problems. Pasteur’s research, which advanced the field of microbiology while helping to cure problems of food spoilage and disease, is the generic example of such work. Stokes proposed replacing the scheme in Figure 4 with a two-dimensional version, shown in Figure 5.

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 5. A 2-D representation of “basic” and “applied” considerations for research (Stokes, 1997, p. 73).

Whereas Whitehurst argued that an emphasis on Edison’s quadrant – in essence, focusing on “What Works” without necessarily asking why – is essential for making progress on curricular issues in mathematics education and other fields, I shall argue that a more balanced approach, with much more significant attention to work in Pasteur’s quadrant, will have a much larger payoff. To do this I shall make a direct analogy with clinical studies in evidence-based medicine. The arguments made here have been influenced by a long-term collaboration with Hugh Burkhardt. An approach parallel to the one discussed here has been called the “engineering approach” to education (see Burkhardt, 2006; Burkhardt & Schoenfeld, 2003).

The U.S. Food and Drug Administration (FDA, 2005) described four “steps from test tube to new drug application review”: preclinical research and three phases of clinical studies (see <http://www.fda.gov/cder/handbook/>). In preclinical drug (or indeed, any medical treatment) research, a wide range of experimentation is done with animals, in order to determine that proposed interventions are likely to be useful and cause no harm to humans. The FDA then defines Phase 1, 2, and 3 studies as follows:

*Phase 1 Clinical Studies.* Phase 1 includes the initial introduction of an investigational new drug into humans. These studies are closely monitored and may be conducted in patients, but are usually conducted in healthy volunteer subjects. These studies are designed to determine the metabolic and pharmacologic actions of the drug in humans, the side effects associated with increasing doses, and, if possible,

to gain early evidence on effectiveness.... The total number of subjects included in Phase 1 studies varies with the drug, but is generally in the range of twenty to eighty. (<http://www.fda.gov/cder/handbook/phase1.htm>, August 14, 2004)

*Phase 2 Clinical Studies.* Phase 2 includes the early controlled clinical studies conducted to obtain some preliminary data on the effectiveness of the drug for a particular indication or indications in patients with the disease or condition. This phase of testing also helps determine the common short-term side effects and risks associated with the drug. Phase 2 studies are typically well-controlled, closely monitored, and conducted in a relatively small number of patients, usually involving several hundred people. (<http://www.fda.gov/cder/handbook/phase2.htm>, August 14, 2004)

*Phase 3 Clinical Studies.* Phase 3 studies are expanded controlled and uncontrolled trials. They are performed after preliminary evidence suggesting effectiveness of the drug has been obtained in Phase 2, and are intended to gather the additional information about effectiveness and safety that is needed to evaluate the overall benefit-risk relationship of the drug. Phase 3 studies also provide an adequate basis for extrapolating the results to the general population and transmitting that information in the physician labeling. Phase 3 studies usually include several hundred to several thousand people. (<http://www.fda.gov/cder/handbook/phase3.htm>, August 14, 2004)

I have abstracted the general approach to phases 1, 2, and 3 on the left-hand side of Figure 6. My argument is that there are analogues of such studies in education, and that it would be useful to follow through with them. The educational analogues are given on the right-hand side of Figure 6.

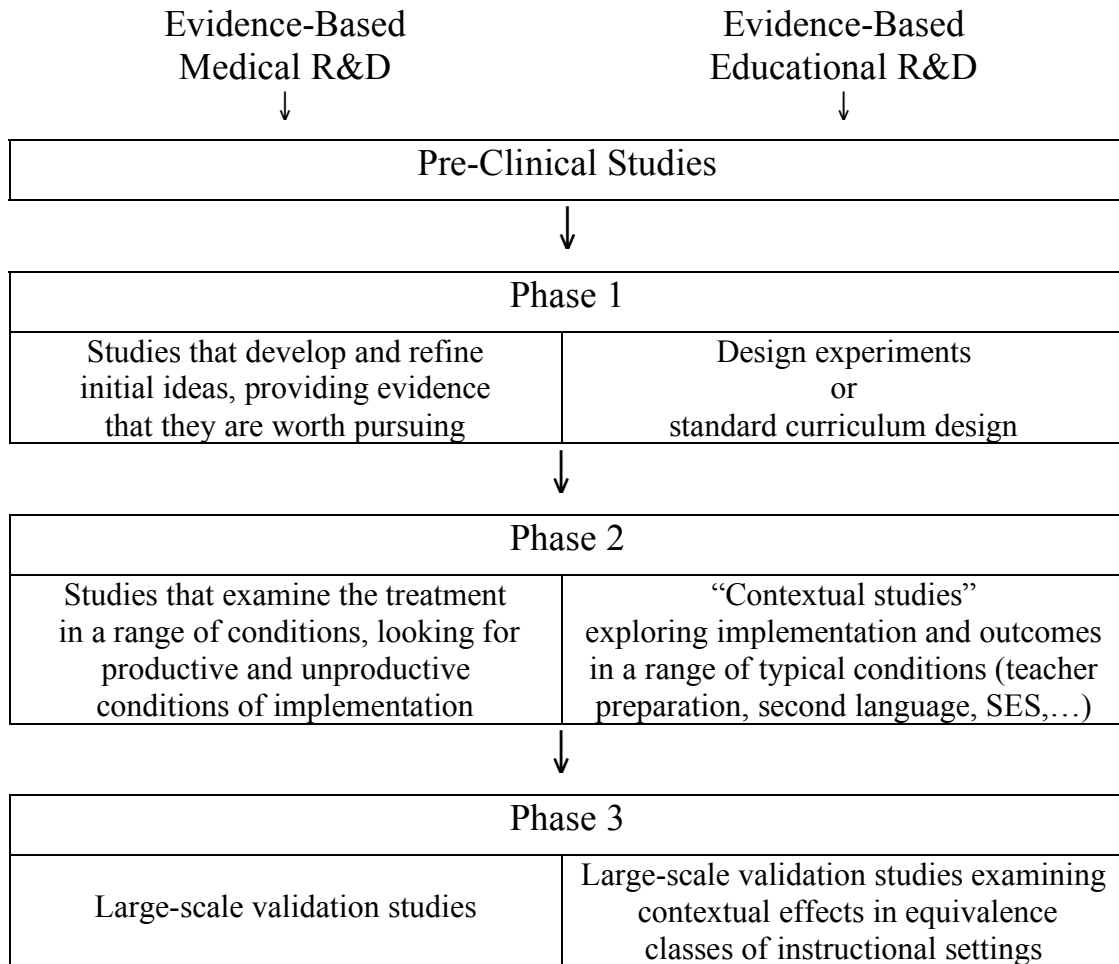


Figure 6. Potential parallels between evidence-based research in medicine and education.

In what follows I flesh out the details of Figure 6.

#### *Preliminary and Phase 1 Studies in Educational R&D*

Research in each of Bohr’s, Edison’s, and Pasteur’s quadrants makes significant contributions as preliminary and Phase 1 studies in educational R&D. Although the emphasis here is on curricular design, it is worth recalling that many curriculum ideas have their origins in Bohr’s quadrant (research with a focus on fundamental understanding) – and in addition that such research encompassed a wide range of methods, from experimental studies in the laboratory to classroom ethnographies. Basic research in the 1970s and 1980s resulted in a fundamental reconceptualization of the nature of mathematical thinking, along multiple dimensions. Theoretical considerations in the field at large resulted in an epistemological shift away from an acquisitionist theory of knowing to a constructivist view, laying the foundation for a different conception of student engagement with mathematics. Laboratory studies revealed the mechanisms by which problem-solving strategies could be elaborated and learned, and demonstrated that poor metacognition could hamper problem solving. Classroom observations and

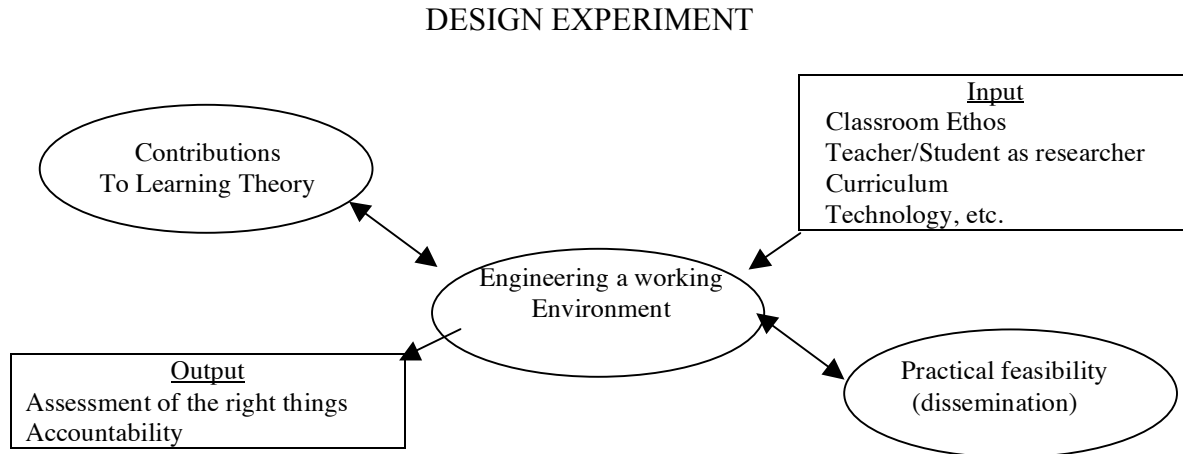
laboratory studies demonstrated the negative impact of counterproductive beliefs on mathematical performance, and classroom studies suggested the origins of such beliefs. All of this work informed NCTM's (1989) *Curriculum and Evaluation Standards for School Mathematics*, which served as a catalyst for the National Science Foundation to issue a series of requests for proposals (RFPs) for curriculum development in the early 1990s. Hence basic research played a fundamental role in shaping curriculum development over the final quarter of the 20th century, and in shaping the criteria by which curricula should be evaluated.

Once the goals and theoretical orientation had been established, fundamentally applied work played a major role in curriculum development. The teams that produced curricula in response to the National Science Foundation's curriculum RFPs differed in character, with some being university-based and some housed in organizations known for the production of instructional materials. However, all shared the property that they drew on whatever sources they could – knowledge from research where available, but equally if not more important given the state of the art, prior experience in curriculum development and, more generally, what has been called the wisdom of practice (see, e.g., Shulman, 2004). Given the timetable for curriculum development and refinement – typically, the grant for the development of an  $n$ -year curriculum had a term of  $n$  years! – the process of curriculum development and refinement necessarily resided squarely in Edison's quadrant.

Over the same time period, work in Pasteur's quadrant – work that simultaneously addressed theoretical issues and pressing issues of practice in fundamental ways – was beginning to become a productive reality. Early exemplars, before the concept of design experiment had been named, were called *apprenticeship environments* (Collins, Brown, & Newman, 1989). These included my (1985a, 1985b) courses in mathematical problem solving, Palincsar and Brown's (1984) work on reciprocal teaching, and Scardamalia and Bereiter's (1985) work on facilitating the writing process. Related concepts within the mathematics education community have included teaching experiments (Cobb, 2000, 2001; Steffe & Thompson, 2000), classroom-based research (Ball, 2000; Ball & Lampert, 1999; Lampert, 2001), and curriculum research as design (Battista & Clements, 2000; Roschelle & Jackiw, 2000).

The term *design experiment* was introduced into the literature by Ann Brown (1992) and Allan Collins (1992). The authors intended for both of the terms in the name to be taken with great seriousness. Design experiments were exercises in instructional *design*: they involved the creation of (novel) instructional environments (contexts and materials) that played out as real instruction for real students. They were also true *experiments* – not necessarily in the statistical sense, though data were often gathered and analyzed statistically, but in the scientific sense, in that theory-based hypotheses were made, measurement tools were established, a deliberate procedure was rigorously employed, outcomes were measured and analyzed, and theory was revisited in light of the data.

Brown (1992) encapsulated this complexity as shown in Figure 7. It is worth noting that the double arrow between “contributions to learning theory” and “engineering a learning environment” represents the dualism inherent in working in Pasteur's quadrant.



*Figure 7.* The complex features of design experiments. (From Brown, 1992, p. 142. With permission)

The meaning of design experiments has not been settled in the literature. Related but clearly distinct formulations may be found in Collins (1999), the Design-Based Research Collaborative (2003), and Cobb, Confrey, diSessa, Lehrer, and Schauble (2003). Collins provided a contrast between traditional psychological methods and the context and methods of typical design experiments. In the latter case one gives up a significant degree of control and straightforward experimental prediction for a much more complex set of human interactions, and the need to develop rich characterizations of those interactions. The Design-Based Research Collaborative noted the interaction of theoretical and empirical goals, and the cyclic nature of design, enactment, analysis, and redesign; they called for methods to “document and connect processes of enactment to outcomes of interest.” Cobb et al. identified a wide range of studies that involve design and fit squarely into Pasteur’s quadrant: small-scale versions of a learning ecology that enable detailed studies of how and why they function the way they do (e.g., Steffe & Thompson, 2000); collaborations between a teacher and a research team to construct, conduct, and assess instruction (e.g., Cobb, 2000, 2001; Confrey & Lachance, 2000; Gravemeijer, 1994); preservice teacher development experiments (Simon, 2000); collaborations between researchers and in-service teachers to support the development of a professional community (Lehrer & Schauble, 2000; Silver, 1996; Stein, Silver, & Smith, 1998); and systemic interventions such as in New York City’s Community School District #2 (High Performance Learning Communities, 2003). They then delineated the following common features of those efforts:

- Purpose: “to develop a class of theories about both the process of learning and the means that are designed to support that learning.”
- “Highly interventionist methodology: Design studies are typically test beds for innovation.”
- Theory testing: “Design experiments create the conditions for developing theories yet must place those theories in harm’s way.”
- “Iterative design.”

- “Pragmatic roots: theories developed during the process of experiment are humble not merely in the sense that they are concerned with domain-specific learning experiences, but also because they are accountable to the activity of design.” (Cobb et al., 2003, pp. 9-10)

Despite the differences in emphasis among those who theorize about design experiments, there is certainly an agreement that one desirable outcome of design experiments is a “road tested” design for an instructional environment or curriculum that shows promise in terms of student learning and does not have any major downside risks.

However one arrives at them – inspired by basic research (origins in Bohr’s quadrant), drafted by skilled designers on the basis of their intuitions or produced by a commercial process (designed or refined in Edison’s quadrant), or produced by a process that attends to both theory and materials development (Pasteur’s quadrant) – the road-tested design described in the previous paragraph has all of the desired properties that a Phase 1 product should have: promise, feasibility, and no obvious risks. So: What next?

One possibility, consistent with much of the current political zeitgeist, is to move directly to randomized controlled trials. After all, if you have a product that you think “works,” why not “prove it,” using what has been called the gold standard of experimental methods? In what follows I shall argue that such an approach, although well intended, represents a mistaken parallel to evidence-based medicine. The next step in clinical trials is to move to Phase 2, mid-scale exploratory studies. There is an appropriate educational analogue, which has not been explored by the field at large.

### *Phase 2 Studies in Educational R&D*

I begin with a story that illustrates the main point underlying the approach suggested in this section. I had been following the use of a standards-based curriculum in local classrooms. Implementation was uneven at times, but what I saw in local middle schools, across the boards, confirmed the impression I had from the literature. In comparison with students who experienced the “traditional” curriculum (represented by two state-adopted textbooks in California), students who experienced the standards-based curriculum performed respectably with regard to skills, and they were better off when it came to concept acquisition and problem solving. At that point, I was leaning toward a blanket recommendation that the newer standards-based curricula be adopted. It appeared that the floor for such curricula was higher than the floor for traditional curricula.

Then I visited an off-track algebra class in a local high school. This was a second semester algebra course being taught in the first semester. All of the students in the class were there because they had had some difficulty with algebra – they were either repeating the course or had entered it from a tracked course that had covered one semester of algebra over the course of a year. By and large, the students were disaffected and saw themselves as poor mathematics learners.

In this school as in many urban schools, privilege played a role in course assignments. The “best” teaching assignments, such as calculus, went to those with the most seniority and clout; the most problematic assignments, of which this was one, went to those who had little choice and sometimes little motivation to teach them. The person

assigned to teach this course was a computer science teacher, and the students were assigned to his room – which had computers bolted to each desk.

The curriculum calls for a great deal of group work. Roughly speaking, students are expected to work 10 problems a day in class, largely in groups; this work prepares them for the 10 problems they are assigned each day for homework. The in-class work is essential, because the homework from this curriculum looks somewhat different from the work in the traditional curriculum. Parents or other caregivers are less likely than usual to be able to help with homework.

Simply put, the class was dysfunctional. The teacher assigned students to groups but provided neither monitoring nor assistance; some students listened to headphones, some drummed basketballs or otherwise indicated their disaffectedness. With computers bolted to their desks, even those students who wanted to work found themselves in an environment that was crowded, noisy, and inhospitable to such work. (The teacher reacted to an early complaint about the level of chaos by saying that the disruptive students would soon drop out, and that the classroom would then be quieter.) Generally speaking, students were either unwilling to work in class or found it nearly impossible to do so. This caused serious problems for those who tried. Unable to work in class, they needed to do 20 problems (the in-class assignment *and* the homework assignment) each night, often without support, in order to keep up.

The fact is that (short of replacing the teacher with a more attentive and competent one, which was not likely to happen) these students would have been better off taking a course that used the traditional curriculum, for the following reasons. Being given straightforward presentations, even those aimed at procedural learning, would have been better than receiving no instruction at all. Seatwork on standard problems might have been doable for a larger percentage of the students. And many of them could have received homework help from parents or caregivers. In sum, in this somewhat pathological context, using a *Standards*-based curriculum actually lowered the floor. In this kind of context the students would not have been well served, but they would have been better served, by the traditional curriculum.

This kind of story will come as no surprise to those who are familiar with urban classrooms. However, it illustrates an essential point that seems to have been missed by many. A curriculum is not a “thing” that is “given” to students, with consistent effects. A curriculum plays out differently in different contexts. The same instructional materials are unlikely to result in the same patterns of learning in inner-city schools that have broken-down facilities and a low proportion of credentialed staff that they produce in suburban schools that have modern facilities and a well-credentialed staff. Some curricula will work well with, or be easily adaptable to, a student body that has a high proportion of English language learners; others, because of their high threshold of formal or academic language, may prove difficult for such students. Some may require a significant amount of teacher preparation before implementation can be successful; some may not. If one is in the position of choosing a curriculum, one wants to be aware of these *contextual effects* and choose a curriculum that has the potential to provide the best results given the real or potential affordances of the present context.

This is precisely the kind of information provided by Phase 2 studies in medical research. The pharmaceutical industry has long understood, for example, that drugs may have differential impacts (some positive, some negative) depending on: the health of the patient and what other conditions he or she may have; on what other medicines the patient may be taking; and on whether the medicine is taken on a full or empty stomach, or with particular foods. Thus, for example, Medline Plus, a web server made available by the U.S. National Library of Medicine and the National Institutes of Health at <http://www.nlm.nih.gov/medlineplus/druginformation.html>, has standard categories of information related to various medical treatments. A typical page describing a drug or treatment regimen includes categories such as the following:

- Before Using This Medicine

This category includes information about allergies, special situations such as pregnancy or breast feeding, appropriateness for particular populations, and interactions with other conditions and/or medicines.

- Proper Use of This Medicine

This category includes information about necessary conditions for proper use. This includes a drug or physical therapy regimen, but also information such as “to be taken on a full stomach” or even as specific as “do not drink grapefruit juice with this medicine.”

- Precautions While Using This Medicine

This category includes descriptions of downside risks to taking the medicine, possible complications, and things that should be avoided (typically, taking alcohol or other medicines that might interact with the one being considered).

In pharmacology, the point of Phase 2 studies is to uncover this kind of information, *before* proceeding to large-scale testing via randomized controlled trials. In broad terms, the issues are, What needs to be in place before the proposed approach stands a decent chance of being effective? Which outcomes should be tracked, in what ways? What makes effectiveness unlikely? What kinds of collateral benefits or costs might one encounter, with what frequency, for what populations? Only after such questions have been answered – when one has a decent sense of where and when something works – should one set about the large-scale documentation *that* it works in those circumstances.

In what follows I explore the educational analogue. Lest I be accused of reductiveness at the start, I should point out that public health may be a better analogy than pharmacology for instruction-in-context. The rough parallel is that curricula play a role in the educational system (where there is room for anything from almost total neglect to faithful implementation of what is proposed) that is analogous to the role that medicines play in public health implementation (where there is room for anything from almost total neglect to faithful implementation of what is proposed). What I am proposing here is somewhat speculative and (like all work in Pasteur’s quadrant) will require a good deal of adaptation and midcourse correction to be fully effective.

Suppose, then, that one has developed some instruction,<sup>16</sup> possibly as part of a larger learning environment. Assume that the instruction extends over a substantial period of time – perhaps a semester, a year, or some number of years. The path toward the development of the instruction may have been any of those described in the section “Preliminary and Phase 1 Studies” – it may have been inspired by research, may have emerged from a team of skilled designers, or may be the product of one or more design experiments. The instruction has presumably been used with a relatively small number of students and found promising. What next?

It seems clear that the instruction should be implemented and examined in a range of contexts that are representative of the wide range of circumstances in which it might be adopted. (These might include suburban schools, inner city schools, schools with low and high proportions of English language learners, schools with low and high proportions of credentialed mathematics teachers, and so on.) This implementation and analysis should be systematized, perhaps as follows.

Although no two schools are the same, it might be possible to identify a collection of “Phase 2 schools” that, in some way, represent equivalence classes of schools. That is, suppose one selects a collection of schools such that each school has a collection of demographic and perhaps administrative attributes that typify a significant number of schools. It is not unreasonable to expect that the implementation in a Phase 2 school that represents an equivalence class would be at least somewhat similar to implementation in other schools in that equivalence class, with somewhat similar results. Hence implementation over a range of schools selected from different equivalence classes would suggest what might happen if the instruction were implemented on a wider scale. It would also indicate which kinds of contexts are likely to experience beneficial results, and which are not.

In each of the Phase 2 schools, one would investigate questions such as the following:

- Under what conditions can the instruction be made to work (and what does one mean by work)?
- What measures, both of process (e.g., classroom dynamics, and discourse patterns) and product (mathematical or other outcomes) can be used or developed to help understand and calibrate the impact of such instruction?
- How does the instruction get operationalized in different contexts, and with different kinds of support systems? What are the patterns of uptake, of interaction?
- How does implementation in different contexts reflect on the theoretical underpinnings of the design?

---

<sup>16</sup> In what follows I use the term *instruction* because it is somewhat generic. I intend a very broad interpretation in what follows. If, for example, the intended instruction requires certain classroom configurations (e.g., the kind of “jigsawing” used by Brown and Campione, 1996, in their FCL work) or certain technology (as in Scardamalia and Bereiter’s, 2002, Knowledge Forum work), those are considered to be part of the instruction.

- How might the curriculum be refined in light of what is revealed by the attempts to implement it?

In the early stages of Phase 2 work in these varied contexts (Phase 2a), the work would resemble a family of coordinated design experiments. Understanding and describing instructional impact in different contexts would still be a contribution to fundamental research at that point; a wide range of methods would still be appropriate to explore the implementation, and to find out what support structures, in different environments, improved outcomes. This kind of work could provide “user’s guides” to the curricula. (If a district has a particular profile, it might expect certain patterns of outcomes. Thus it might want to consider modifying certain administrative features of the environment and providing professional development along certain lines to its staff.)

One would, of course, expect some aspects of this kind of work to be more efficient than in a first design experiment: Some of the key phenomena would have been noted in prior R&D, and some of the student assessment tools would have been developed. Indeed, in the interests of efficiency and with an eye toward later studies, one would want to have or develop similar observational and assessment tools for as many of these contexts as possible.

Once the basic research in the different contexts has been done, efforts turn toward systematization (Phase 2b). Are there features of implementation fidelity that are essential to produce certain kinds of student outcomes? Looking across the various contexts in which the instruction has been implemented, are there features of those contexts that seem consistently to produce certain kinds of effects? Are there certain contextual variables that, having been identified in the more detailed studies, can be captured more efficiently in summary fashion once one knows to look for them? (For example, early studies of accountability – to what and to whom are students accountable, and how does that play out in terms of learning and identity? – can be expected to be highly descriptive and detailed. After some time, however, one can imagine that the character of accountability structures could be codified in straightforward ways.)

A major goal of the work in Phase 2b is to develop the instrumentation that supports the meaningful conduct of Phase 3 studies. This includes the ability to capture central aspects of the context as described in the previous paragraph. It includes broader contextual variables found to have an impact on implementation or outcomes. (These might, for example, include larger aspects of the context, such as whether the state or district has high-stakes accountability tests of particular types. Given that many teachers and schools depart from planned curricula for substantial periods of the school year in order to focus on test preparation, the presence of such tests might have a significant impact and should be included as a variable.) It includes characterizations of implementation fidelity. And it includes a range of outcome measures. Content-related measures would, in the best of circumstances, be independent of the curriculum and standards-based. (Recall that a major problem with early curricular evaluations such as those found in Senk and Thompson, 2003, is that many of the student assessments were locally developed, and thus not comparable to other such measures.) Other measures might well include characterizations of the learning community, of student beliefs, and of other aspects of the students’ mathematical identities. Some comparative studies would be conducted during either Phase 2a or 2b, in order to assess the relative impact of the

instruction and gauge the plausibility of the claim that the instruction is indeed worth adopting. To be explicit in summary, much of Phase 2 research as imagined here would reside squarely in Pasteur's quadrant – contributing to basic understandings, but also laying the foundations for the kind of applied work done in Phase 3.

*Phase 3 Studies in Educational R&D.*

The main focus of Phase 2 was to understand how and why instruction “works,” to revise and improve that instruction, to identify relevant variables, and to develop relevant instrumentation. With that established, Phase 3 can be both straightforward and informative. One can imagine two kinds of Phase 3 studies: (1) expanded studies replicating the effects discovered in Phase 2, and (2) comparative in-depth evaluations of any number of promising curricula. The two kinds of studies can be done independently, or studies of type 1 (which are in essence comparative studies of two treatments, the instruction and a control) can be done as subsets of studies of type 2.<sup>17</sup>

A key aspect of such studies is the random assignment of students to instructional treatments. That said, the rest of the work should be reasonably straightforward, in conceptual terms.<sup>18</sup> That is, it should be, *if* the proper work has been done in Phase 2, and the right perspective is taken toward the analysis of the data. As noted above, major goals for research in Phase 2 are to identify contextual variables and be able to code for them, and to use appropriately independent and meaningful assessments of student performance. With such tools at one's disposal, the relevant issue becomes one of interpretation.

Recall the second example discussed in Part 1, which illustrated two different ways to think about the results of randomized controlled trials. In this author's opinion, aggregating across all contexts – “on average, students who studied from Curriculum A outperformed students who studied from Curriculum B” – is of little value. Such an approach provides evaluative information in summary, but it provides little useful information for anyone faced with a choice of instructional materials. (What if, for example, Curriculum A outperformed Curriculum B in general, but Curriculum B outperformed Curriculum A by a significant margin in schools that are similar to the ones for which instructional materials must be chosen? In that case Curriculum B is the right choice, even though Curriculum A is better on average. Or, what if the research reveals that Curriculum B is superior under present circumstances, but that a program of professional development will equip your staff to employ Curriculum A to better effect?) In sum, the most beneficial role of randomized controlled trials in curricular evaluation would be to investigate the hypotheses raised in Phase 2 studies – to determine which contextual variables have which effects on implementation fidelity of a wide range of curricula, and what the effects of those curricula are in those conditions. That kind of information would be truly useful for decision makers.

---

<sup>17</sup> For a broad discussion of comparative studies, as part of the “engineering approach,” see Burkhardt (2006). For an extended treatment of some of the relevant statistical issues, see Scheaffer (2006).

<sup>18</sup> Of course, the devil is in the details. To say statistical analyses are conceptually straightforward is one thing; to carry them out correctly can be something else altogether. Even issues of the appropriate unit of analysis are non-trivial.

## **Coda**

I began doing research in mathematics education in the mid-1970s. At the time, the field's primary research methods (in the United States, at least) were statistical, but their use was often unsophisticated, and the field as a whole suffered somewhat from a reductive form of what has been called "science envy." The field of cognitive science had not yet coalesced (the first annual meeting of the Cognitive Science Society was held in 1978), and many of the fields that would ultimately contribute both methods and perspectives to our current understanding of mathematical thinking and learning (among them anthropology, artificial intelligence, linguistics, philosophy, psychology, and sociology) were at best peripheral to the enterprise. When I began my research on problem solving, the cutting edge of research consisted of attempts to make sense of factors that shaped the success or failure of individuals working problems in isolation in the laboratory. Major constructs that have now come to be cornerstones of our understanding (the very notion of cognitive modeling; the roles of metacognition and beliefs; the concepts of identity and of communities of practice) and the methods that help to elaborate them were only beginning to emerge, if they were on the horizon at all.

The progress that has been made since that time is nothing short of phenomenal. In little more than a quarter century there have been major epistemological shifts, accompanied by a flourishing of tools, techniques, and the theoretical perspectives that underlie them. The cognitive sciences and sociocultural research within mathematics education have both matured and become increasingly robust; what at first seemed almost like thesis and antithesis have, over the past decade or so, begun a synthesis that seems increasingly promising in terms of its possibility to help explain issues of (mathematical) thinking, teaching, and learning. The same is the case for the artificial distinction between quantitative and qualitative methods, which becomes less and less important as one begins to ask these central research questions: "What assumptions are being made? What claims are being made? What warrants are being offered for those claims?" This is remarkable progress, and one hopes and expects to see more.

## **Acknowledgments**

I am most grateful to Jim Greeno, Randi Engle, Tom Carpenter, Natasha Speer, Ellen Lagemann, Andreas Stylianides, Cathy Kessel, Dor Abrahamson, and proximal and distal members of the Functions Group including Markku Hannula, Mara Landers, Mari Levin, Katherine Lewis and Daniel Wolfroot for their penetrating and thought-provoking comments on earlier drafts of this chapter. Frank Lester has been an ideal editor, demonstrating remarkable patience and attention to detail. The final version of this chapter is much improved as a result of their help.

## REFERENCES

- 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.
- Ausubel, D.P. (1960). The use of advance organizers in the learning and retention of meaningful verbal material. *Journal of Educational Psychology*, 51, 267-272.
- Ball, D. (1988). *Knowledge and reasoning in mathematical pedagogy: Examining what prospective teachers bring to teacher education*. Unpublished doctoral dissertation, Michigan State University.
- Ball, D. (2000). Working on the inside: Using one's own practice as a site for studying teaching and learning. In A. E. Kelley & R. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 365-402). Mahwah, NJ: Erlbaum.
- 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.
- Battista, M., & Clements, D. (2000). Mathematics curriculum development as a scientific endeavor. In A. E. Kelley & R. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 737-760). Mahwah, NJ: Erlbaum.
- Bauersfeld, H. (1980). Hidden dimensions in the so-called reality of a mathematics classroom. *Educational studies in mathematics*, 11(1), 109-136.
- Bhattachargee, Y. (2005). Can randomized trials answer the question of what works? *Science*, 25, 1861-1862.
- Boaler, Jo. (2002) *Experiencing school mathematics* (Rev. ed.). Mahwah, NJ: Erlbaum.
- Boaler, Jo. (in press). Promoting relational equity in mathematics classrooms – Important teaching practices and their impact on student learning. Text of a 'regular lecture' given at the 10th International Congress of Mathematics Education (ICME X), 2004, Copenhagen. To appear in the *Proceedings of the 10th International Congress of Mathematics Education*.
- Brousseau, G. (1997). *Theory of didactical situations in mathematics: Didactique des mathématiques 1970-1990*. (N. Balacheff, M. Cooper, R. Sutherland, & V. Warfield, Eds. & Trans.). Dordrecht, The Netherlands: Kluwer.
- Brown, A. (1987). Metacognition, executive control, self-regulation, and other more mysterious mechanisms. In F. Reiner & R. Kluwe (Eds.), *Metacognition, motivation, and understanding* (pp. 65-116). Hillsdale, NJ: Erlbaum.
- 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(2), 1978, 155-192.
- Brown, J.S., & VanLehn, K. (1982). Toward a generative theory of bugs. In T. Carpenter, J. Moser, & T. Romberg (Eds), *Addition and subtraction: A cognitive perspective* (pp. 117-135). Hillsdale, NJ: Erlbaum.
- Brownell, W. (1947). An experiment on "borrowing" in third-grade arithmetic. *Journal of Educational Research*, 41(3), 161-171.
- Bruning, J.L., and Kintz, B.L. (1987). *Computational Handbook of Statistics*. (Third edition). Scott, Foresman.
- Burkhardt, H. (2006). From design research to large-scale impact: Engineering research in education. In J.Akker, K. Gravemeijer, S. McKenney, & N. Nieveen (Eds.), *Design and development research: Emerging perspectives*. Manuscript in preparation.
- Burkhardt, H., & Schoenfeld, A. H. (2003). Improving educational research: Toward a more useful, more influential, and better funded enterprise. *Educational Researcher* 32(9), 3-14.
- Carpenter, T. P., Fennema, E., & Franke, M. L. (1996). Cognitively guided instruction: A knowledge base for reform in primary mathematics instruction. *Elementary School Journal*, 97(1), 1-20.
- Carpenter, T. P., Lindquist, M. M., Matthews, W., & Silver, E. A. (1983). Results of the third NAEP mathematics assessment: Secondary school. *Mathematics Teacher*, 76(9), 652-659.
- Chappelle, M. (2003). Keeping mathematics front and center: Reaction to middle-grades curriculum projects research. In S. Senk & D. Thompson (Eds.), *Standards-based school mathematics curricula: What are they? What do students learn?* (pp. 285-298). Mahwah, NJ: Erlbaum.
- 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.
- Cobb, P. (2000). Conducting teaching experiments in collaboration with teachers. In A. E. Kelly & R. A. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 307-333). Mahwah, NJ: Erlbaum.
- Cobb, P. (2001). Supporting the improvement of learning and teaching in social and institutional context. In S. M. Carver & D. Klahr (Eds.), *Cognition and instruction: Twenty-five years of progress* (pp. 455-478). Mahwah, NJ: Erlbaum.
- Cobb, P., Confrey, J. diSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher*, 32(1), 9-13.

- Cobb, P., & Hodge, L. L. (2002). A relational perspective on issues of cultural diversity and equity as they play out in the mathematics classroom. *Mathematical Thinking and Learning*, 4(2&3), 249-284.
- Cohen, D. (1990). A revolution in one classroom: The case of Mrs. Oublier. *Education Evaluation and Policy Analysis*, 12(3), 311-329.
- Collins, A. (1992). Toward a design science of education. In E. Scanlon & T. O'Shea (Eds.), *New directions in educational technology* (pp. 15-22). Berlin, Germany: Springer.
- Collins, A. (1999). The changing infrastructure of educational research. In E. Lagemann & L. Shulman (Eds.), *Issues in education research: Problems and possibilities* (pp. 289-298). New York: Jossey-Bass.
- Collins, A., Brown, J.S., & Newman, S. (1989). Cognitive apprenticeship: Teaching the craft of reading, writing, and mathematics. In L. B. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honor of Robert Glaser* (pp. 453-494). Hillsdale, NJ: Erlbaum.
- Confrey, J., & Lachance, A. (2000). Transformative reading experiments through conjecture-driven research design. In A. E. Kelly & A. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 231-266). Mahwah, NJ: Erlbaum.
- Conrad, C., & Serlin, R. (Eds.) (2005) *The Sage Handbook for Research in Education: Engaging Ideas and Enriching Inquiry*. Thousand Oaks, CA: Sage.
- Cooney, T. (1985). A beginning teacher's view of problem solving. *Journal for Research in Mathematics Education*, 16, 324-336.
- deCorte, E., Greer, B., & Verschaffel, L. (1996). Mathematics teaching and learning. In D. Berliner & R. Calfee (Eds.), *Handbook of Educational Psychology*, 491-549. New York: MacMillan.
- Denzin, N.K., and Lincoln, Y.S. (Eds.) (2005) *The Sage Handbook of Qualitative Research*. Thousand Oaks, CA: Sage.
- Design-Based Research Collective. (2003) Design-based research: an emerging paradigm for educational inquiry. *Educational Researcher*, 32(1), 5-8.
- diSessa, A. (1983). Phenomenology and the evolution of intuition. In D. Gentner & A. Stevens (Eds.), *Mental models* (pp. 15-34). Hillsdale, NJ: Erlbaum.
- 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.
- Engle, R., & Conant, F. (2002) Guiding principles for fostering productive disciplinary engagement: Explaining emerging argument in a community of learners classroom. *Cognition and Instruction*, 20(4), 399-483.

- Ericsson, K.A., & Simon, H.A. (1980). Verbal reports as data. *Psychological Review*, 87(3), 215-251.
- Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.
- Fawcett, H. (1938). *The nature of proof*. (Thirteenth Yearbook of the National Council of Teachers of Mathematics.) New York: Teachers College.
- Gardner, H. E. (1985). *The mind's new science : A history of the cognitive revolution*. New York: Basic Books.
- Geertz, C. (1975). On the nature of anthropological understanding. *American Scientist*, 63, 47-53.
- Gravemeijer, K. (1994). Educational development and developmental research. *Journal for Research in Mathematics Education*, 25, 443-471.
- Green, J. L., Camilli, G., & Elmore, P. B. (Eds.) (2006) *Handbook of Complementary Methods in Education Research*. Mahwah, NJ: Erlbaum.
- Gutstein, E. (2003). Teaching and learning mathematics for social justice in an urban, Latino school. *Journal for Research in Mathematics Education*, 34, 37-73.
- Hatano, G., & Inagaki, K. (1991). Sharing cognition through collective comprehension activity. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 331–348). Washington, DC: American Psychological Association.
- High performance learning communities project (HPLC). (n.d.). Retrieved August 9, 2003, from <http://www.lrdc.pitt.edu/hplc/>.
- Inagaki, K. (1981). Facilitation of knowledge integration through classroom discussion. *Quarterly Newsletter of the Laboratory of Comparative Human Cognition*, 3(2), 26–28.
- Inagaki, K., Hatano, G., & Morita, E. (1998). Construction of mathematical knowledge through whole-class discussion. *Learning and Instruction*, 8, 503–526.
- Kerstyn, C. (2001). *Evaluation of the I CAN Learn® mathematics classroom: First year of implementation (2000–2001 school year)*. (Available from the Division of Instruction, Hillsborough County Public Schools, Tampa, FL)
- Keeves, J. (Ed.) (1997) *Educational research, methodology and measurement: An international handbook*. Amsterdam: Elsevier.
- 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. Hatfield (Ed.), *Mathematical problem solving* (pp. 7-20). Columbus, OH: ERIC.
- Lampert, M. (1990). When the problem is not the question and the solution is not the answer. *American Educational Research Journal*, 27(1), 29-63.

- Lampert, M. (2001). *Teaching problems and the problems of teaching*. New Haven: Yale University Press.
- LeCompte, M., Millroy, W., & Preissle, J. (Eds.). (1992). *Handbook of qualitative research in education*. New York: Academic Press.
- Lee, J. (2002). Racial and ethnic achievement gap trends: Reversing the progress toward equity? *Educational Researcher*, 31(1), 3-12.
- Lehrer, R., & Schauble, L. (2000). Modeling in mathematics and science. In R. Glaser (Ed.), *Advances in instructional psychology: Educational design and cognitive science* (pp. 101-159). Mahwah, NJ: Erlbaum.
- Leinhardt, G. (1990). A contrast of novice and expert competence in math lessons. In J. Lowyck, & C. Clark (Eds.), *Teacher thinking and professional action* (pp. 75-97). Leuven, Belgium: Leuven University Press.
- Leinhardt, G., Putnam, R. T., Stein, M. K., & Baxter, J. (1991). Where subject knowledge matters. In J. Brophy (Ed.), *Advances in research on teaching* (Vol. 2, pp. 87-113). Greenwich CT: JAI Press.
- Lemke, M., Sen, A., Pahlke, E., Partelow, L., Miller, D., Williams, T., Kastberg, D., Jocelyn, L. (2004). *International outcomes of learning in mathematics literacy and problem solving: PISA 2003 results from the U.S. Perspective*. (NCES 2005-003). Washington, DC: U.S. Department of Education, National Center for Education Statistics.
- Lester, F. (1994). Musings about mathematical problem-solving research: 1970-1994. *Journal for Research in Mathematics Education*, 25(6), 660-675.
- Lucas, J., Branca, N., Goldberg, D., Kantowsky, M., Kellogg, H., & Smith, J. (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. 345-378. Columbus, OH: ERIC
- Mayer, R. (1985). Implications of cognitive psychology for instruction in mathematical problem solving. In E.A. Silver (Ed.), *Learning and teaching mathematical problem solving: Multiple research perspectives* (pp. 123-138). Hillsdale, NJ: Erlbaum.
- Medline Plus. (2005). Hormone replacement therapy. Retrieved May 15, 2005, from <http://www.nlm.nih.gov/medlineplus/hormonereplacementtherapy.html>
- Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review*, 63, 81-97.
- Moll, L. C., Amanti, C., Neff, D., & Gonzalez, N. (1992). Funds of knowledge for teaching: Using a qualitative approach to connect homes and classrooms. *Theory Into Practice*, 31(2), 132-141.
- Muldoon, M.F., Manuck, S.B., and Matthews, K.A. 1990. Lowering cholesterol concentrations and mortality: A quantitative review of primary prevention trials. *British Medical Journal* 301, 309-314.

- Mullis, I. Martin, M., Beaton, A., Gonzalez E., Kelly, D., & Smith, T. (1998). *Mathematics and science achievement in the final year of secondary school*. Boston: The International Association for the Evaluation of Educational Achievement, at Boston College.
- Mullis, I., Martin, M., Gonzalez, E., Gregory, K., Garden, R., O'Connor, K., Chrostowski, S., & Smith, T. (2000). *TIMSS 1999: Findings from IEA's repeat of the third international mathematics and science study at the eighth grade. International mathematics report*. Boston: The International Association for the Evaluation of Educational Achievement, at Boston College.
- National Academy of Education. (1999). *Recommendations Regarding Research Priorities: An Advisory Report to the National Educational Research Policy and Priorities Board*. Washington, DC: National Academy of Education. Available for download from <http://www.nae.nyu.edu/pubs/index.htm>.
- National Council of Teachers of Mathematics. (1989) *Curriculum and evaluation standards for school mathematics*. Reston, VA: NCTM.
- National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: NCTM.
- National Research Council. (1994). *Women and health research: Ethical and legal issues of including women in clinical studies, Vol. 1*. (A. C. Mastroianni, R. Faden, and D. Federman, Editors). Washington, DC: Institute of Medicine, National Academy Press.
- National Research Council. (2001). *Adding it up: Helping children learn mathematics*. Washington DC: National Academy Press.
- National Science Foundation. (1997). *Foundations: The challenge and promise of K-8 science education reform, Vol. 1 (NSF 97-76)*. Washington, DC: Author.
- Orne, Martin T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, 17(11), 776-783.
- Palincsar, A. S., & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction*, 2, 117-175.
- Pickering, A. (1995). *The Mangle of Practice: Time, Agency, and Science*. Chicago: University of Chicago Press.
- Popper, K. (1963). *Conjectures and Refutations: The Growth of Scientific Knowledge*. London: Routledge, Keagan & Paul.
- Putnam, R. (2003). Commentary on four elementary mathematics curricula. In S. Senk & D. Thompson (Eds.), *Standards-based school mathematics curricula: What are they? What do students learn?* (pp. 161-178). Mahwah, NJ: Erlbaum.
- Ridgway, J., Crust, R., Burkhardt, H., Wilcox, S., Fisher, L., & Foster, D. (2000). *MARS report on the 2000 tests*. San Jose, CA: Mathematics Assessment Collaborative.

- Riley, J. (1990). *Getting the most from your data: A handbook of practical ideas on how to analyse qualitative data*. Bristol, England: Technical and Educational Services Ltd.
- Roschelle, J., & Jackiw, N. (2000). Technology design as educational research: Interweaving imagination, inquiry, and impact. In A. E. Kelley & R. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 777-798). Mahwah, NJ: Erlbaum.
- Rosebery, A., Warren, B., & Conant, F. (1992). Appropriating scientific discourse: Findings from language minority classrooms. *Journal of the Learning Sciences*, 2, 61-94.
- Rosnick, P., & Clement, C. (1980). Learning without understanding: the effect of tutoring strategies on algebra misconceptions. *Journal of Mathematical Behavior*, 3(1), 3-27.
- Saxe, G., & Schoenfeld, A. (1998). Annual meeting, 1999. *Educational Researcher*, 27(5), 33.
- Scardamalia, M. (2002). Collective cognitive responsibility for the advancement of knowledge. In B. Smith (Ed.), *Liberal education in a knowledge society* (pp. 67-98). Chicago: Open Court.
- Scardamalia, M., & Bereiter, C. (1985). Fostering the development of self-regulation in children's knowledge processing. In S. F. Chipman, J. W. Segal, & R. Glaser (Eds.), *Thinking and learning skills: Research and open questions* (pp. 563-577). Hillsdale, NJ: Erlbaum.
- Scardamalia, M., & Bereiter, C. (1991). Higher levels of agency for children in knowledge building: A challenge for the design of new knowledge media. *Journal of the Learning Sciences*, 1, 37-68.
- Scardamalia, M., Bereiter, C., & Lamon, M. (1994). The CSILE project: Trying to bring the classroom into World 3. In K. McGilley, ed., *Classroom Lessons: Integrating Cognitive Theory and Classroom Practice* (pp. 201-228). Cambridge, MA: MIT Press.
- Scheaffer, Richard L. (Ed.) (2006). *Guidelines for reporting and evaluating mathematics education research*. Proceedings of an NSF-Supported workshop series on the uses of statistics in mathematics education research. Manuscript in preparation.
- Schoenfeld, A. H. (1983). Beyond the purely cognitive: Belief systems, social cognitions, and metacognitions as driving focuses in intellectual performance. *Cognitive Science*, 7, 329-363.
- Schoenfeld, A. (1985a). *Mathematical problem solving*. New York: Academic Press.
- Schoenfeld, A. (1985b). Metacognitive and epistemological issues in mathematical understanding. In E. A. Silver (Ed.), *Teaching and learning mathematical problem solving: Multiple research perspectives* (pp. 361-380). Hillsdale, NJ: Erlbaum.
- Schoenfeld, A. H. (1988) When good teaching leads to bad results: The disasters of well taught mathematics classes. *Educational Psychologist*, 23 (2), 145-166.

- Schoenfeld, A. H. (1992). Learning to think mathematically: Problem solving, metacognition, and sense-making in mathematics. (1992). In D. Grouws (Ed.), *Handbook for Research on Mathematics Teaching and Learning*, pp. 334-370. New York: MacMillan.
- Schoenfeld, A. H. (1998). Toward a theory of teaching-in-context. *Issues in Education*, 4(1), 1-94.
- Schoenfeld, A. H. (2002). 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. (2004). The math wars. *Educational Policy*, 18(1), 253-286.
- Schoenfeld, A. H. (2006a). 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(2), 13-21.
- Schoenfeld, A. H. (2006b). Reply to Comments From the What Works Clearinghouse on "What Doesn't Work." *Educational Researcher*, 35(2), 23.
- Senk, S., & Thompson, D. (Eds.) (2003). *Standards-based school mathematics curricula: What are they? What do students learn?* Mahwah, NJ: Erlbaum.
- Shirk, G. B. (1972). *An examination of the conceptual frameworks of beginning mathematical teachers*. Unpublished dissertation. Urbana-Champaign: University of Illinois.
- Shulman, L. S. *The Wisdom of Practice: Essays on Teaching, Learning, and Learning to Teach*. San Francisco: Jossey-Bass.
- Siegler, R. S. (1987). The perils of averaging data over strategies: An example from children's addition. *Journal of Experimental Psychology: General*, 116, 250-264.
- Silver, E. (1996). Moving beyond learning alone and in silence: observations from the QUASAR project concerning communication in mathematics classrooms. In L. Schauble & R. Glaser (Eds.), *Innovations in learning: New environments for education* (pp. 127-159). Mahwah, NJ: Erlbaum.
- Simon, M. A. (2000). Research on the development of mathematics teachers: The teacher development experiment. In A. E. Kelly & R. A. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 335-359). Mahwah, NJ: Erlbaum.
- Smith, J., diSessa, A. & Roschelle, J. (1993/1994). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *The Journal of the Learning Sciences*, 3, 115-163.
- Sowder, L. (1985). Cognitive psychology and mathematical problem solving: A discussion of Mayer's paper. In E.A. Silver (Ed.), *Learning and teaching mathematical problem solving: Multiple research perspectives* (pp. 139-146). Hillsdale, 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.
- Stein, M. K., Silver, E. A., & Smith, M. S. (1998). Mathematics reform and teacher development: A community of practice perspective. In J. G. Greeno & S. V. Goldman (Eds.), *Thinking practices in mathematics and science learning* (pp. 17-52). Mahwah, NJ: Erlbaum.
- Stigler, J., & Hiebert, J. (1999). *The teaching gap*. New York: Free Press.
- Stodolsky, S. S. (1985). Telling math: Origins of math aversion and anxiety. *Educational Psychologist* 20, 125-133.
- Stokes, D. E. (1997). *Pasteur's quadrant: Basic science and technical innovation*. Washington, DC: Brookings.
- Swafford, J. (2003). Reaction to high school curriculum projects' research. In Sharon Senk & Denise Thompson (Eds.), *Standards-based school mathematics curricula: What are they? What do students learn?* (pp. 457-468). Mahwah, NJ: Erlbaum.
- Tashakkori, A., & Teddlie, C. (Eds.). (2002). *Handbook of Mixed Methods in Social & Behavioral Research*. Thousand Oaks, CA: Sage.
- Toulmin, S. (1958). *The Uses of Argument*. Cambridge: Cambridge University Press.
- U.S. Department of Education (2002). *Strategic Plan*. Washington, DC: U.S. Department of Education (2002).
- U.S. Food and Drug Administration. (2004) Definitions of Phase 1, Phase 2, and Phase 3 clinical research. Downloaded from <http://www.fda.gov/cder/handbook/phase1.htm>, <http://www.fda.gov/cder/handbook/phase2.htm>, and <http://www.fda.gov/cder/handbook/phase3.htm>, August 14, 2005.
- VanLehn, K., Brown, J.S., & Greeno, J.G. (1984). Competitive argumentation in computational theories of cognition, In W. Kintsch, J. R. Miller and P. G. Polson (Eds.), *Methods and tactics in cognitive science* (pp. 235-262). Hillsdale, NJ: Erlbaum.
- What Works Clearinghouse. (2004). Detailed study report: Kerstyn, C. (2001). Evaluation of the I CAN LEARN Mathematics Classroom. First year of implementation (2000–2001 school year). Unpublished manuscript. Downloaded from <http://www.whatworks.ed.gov/>, February 27, 2005
- Whitehurst, G. 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. Available at <http://www.ed.gov/rschstat/research/pubs/ies.html>
- Whitehurst, G. (2003). Evidence-based education. Powerpoint presentation dated June 9, 2003. Downloaded from <http://www.ed.gov> on November 25, 2005.

- World Health Organization. (1998). Gender and health: Technical paper 98.16 (Reference WHO/FRH/WHD/98.16) New York: World Health Organization.
- Wysowski, D.K., Kennedy, D.L., and Gross, T.P. 1990. Prescribed use of cholesterol-lowering drugs in the United States, 1978 through 1988. *Journal of the American Medical Association* 263(16), 2185-2188
- Yackel, E., & Cobb, P. (1996). Sociomathematical norms, argumentation, and autonomy in mathematics. *Journal for Research in Mathematics Education*, 27(4), 458-477.