

**Learning from Attempts to Improve Schooling:  
The Contribution of Methodological Diversity**

By Stephen W. Raudenbush  
University of Michigan

December 2, 2004

NOTE: This paper was commissioned as background for the December 14, 2004 forum, “Applying Multiple Social Science Research Methods to Educational Problems”. The forum was convened by the Center for Education of the National Research Council, with support from the American Educational Research Association, the American Psychological Association, the Decade of Behavior, and the National Science Foundation. Opinions and statements included in the paper are solely those of the individual author, and are not necessarily adopted or endorsed or verified as accurate by the Center for Education or the National Academy of Sciences, including the National Academy of Engineering, Institute of Medicine, or National Research Council.

## Abstract

Educational research is an interdisciplinary effort long characterized by methodological diversity. I illustrate how integration of methodological approaches has produced key insights. I then ask: Why do we hear an urgent call for mixed methods now? Apparently, a recent shift in the applied research agenda, emphasizing the identification of the causal effects of instructional interventions, has fostered concern that methodological pluralism is at risk. In this paper, I argue a) that a focus on evaluating the impacts of interventions designed to improve teaching and learning is entirely appropriate in light of current policy dilemmas; b) that randomized experiments *are* the gold standard for assessing these effects; but that c) the success of the enterprise depends on a well-integrated, methodologically diverse research effort. This effort must precisely define viable instructional aims and outcome variables; identify, refine, and test promising new interventions; and clarify the instructional needs of children in various settings. It must learn not only whether interventions work but also how they work and for whom. I sketch how diverse methods might be combined, and how a healthy scientific community might collaborate to achieve these aims and to generate adequate funding to support this vital enterprise.

## **I. What Are Mixed Methods and Why Are We Talking About Them Now?**

*Education* may be regarded broadly as the process by which people are taught the language, ideas, meanings, expectations, and knowledge they need to interact successfully in society; or more narrowly, as the formal institutional process occurring in all modern societies that assigns students to classrooms supervised by teachers in schools to learn more specialized aspects of the culture. Either way, the study of education is of central interest to all social science disciplines, so it is not surprising that researchers have long used the enormous array of methodological approaches characteristic of these disciplines as they study education. Against this background of disciplinary and methodological diversity, an outside observer might be puzzled to hear current call for the use of “mixed” or “multiple” methods in educational research. What does this call mean and why do we hear it now?

### **What Do We Mean by “Mixed-Methods Research?”**

It is useful to distinguish *ex ante* versus *post hoc* strategies for inquiry using multiple methods. Using the *ex ante* approach, one deliberately designs new data collection to combine varied methodological approaches in a way that will capitalize on the strengths of each approach. Using the *post hoc* approach, one synthesizes past research, taking care to insure that the studies to be synthesized reflect varied methodological approaches. Either way, the implication is that “mixed methods research” goes beyond a laissez-faire tolerance for the diversity of methods used by varied disciplines. Instead, the idea is to adopt a self-conscious and explicitly articulated strategy for combining information from multiple methods based on a belief that the resulting

inquiry will be more credible, more useful, or more comprehensive than would have been the case if any single methodological approach had been adopted alone.

What counts for methodological diversity? When is an inquiry a “mixed-methods” inquiry, and when is it not? Many discussions of mixed-methods research (see Johnson and Onwuegbuzie, 2004) emphasize the *quantitative* versus *qualitative* dimension as key. It appears that for an inquiry to be anointed as a “mixed-method” inquiry, it must involve both quantitative and qualitative data. I find no logical requirement that the quantitative-qualitative distinction be privileged as the key source of methodological variation in educational research. Quantitative and qualitative research are each internally diverse, and one might argue that surveys, which attempt to describe the social world without changing it, have more in common with qualitative interview studies than with experiments; experiments, unlike surveys or qualitative interview studies, attempt to learn about the world by deliberately intervening to change it.

The privileged status of the qualitative-quantitative polarity presumably reflects the longstanding “paradigm war” within educational research, typically characterized as the battle between quantitative versus qualitative paradigms, though some use the terms “nomothetic versus idiographic” or, more pejoratively, “positivistic versus naturalistic” (Lincoln and Guba, 1985) to characterize this distinction. In any case, mixed-methods research has been advocated as a peaceful resolution to that war and even as a third paradigm in itself (Johnson and Onwuegbuzie, 2004)!

While I have never been comfortable with the characterization of these perspectives in paradigmatic terms, there are useful distinctions between the kinds of understanding their practitioners most often pursue, and a good case can be made that the

varied kinds of understanding that emerge can be complementary, each enriching the other. These distinctions arise most interestingly for me in examples rather than in attempts to characterize epistemological differences. Following Phillips (1983), I find attempts among educational researchers to equate methodological differences with epistemological differences confusing.

Notwithstanding its asserted status as a new, third paradigm, the deliberate blending of quantitative and qualitative data has a long history in social science. In part, this reflects the unique persuasive appeal of different kinds of evidence. Qualitative accounts, for example, often create vivid imagery regarding conditions that need to be rectified while quantitative evidence confirms that those conditions are widespread. National action is therefore required. Thus, for example, in portraying the ravages of early capitalism in England, Engels (1844) combined statistical evidence on life expectancy, housing conditions, and household income of working people with detailed and deeply disturbing observational accounts of life in Manchester, Glasgow, and other British cities at the dawn of the industrial revolution. This work can readily be regarded as “mixed-methods research,” albeit from a polemical perspective only rarely encountered in today’s educational research.

Historical research might generally be regarded as qualitative in method, typically using documents and artifacts to construct narrative accounts. Yet historians are happy to make use of quantitative data and will undoubtedly do so with increased frequency given the tendency of modern societies to preserve large archives of data. In his account of the role of public schooling in the “Americanization” of European immigrants in Chicago, Mirel (2004) relies primarily on articles and letters in local newspapers published by

German, Polish, Bohemian, Czech, and Hungarian immigrant groups, creating a narrative account of how these groups understood and advocated patriotic Americanism. Contrary to popular belief, these groups tended to conceive their pride in America as fully consistent with their pride in the language and history of the motherland, and in fact tended to regard their freedom to maintain their own language, culture, and religion as a distinctly American virtue. To set the context for this local portrait, Mirel relied on national statistics on rates of immigration by year and by country. He also invoked evidence on public and parochial school enrollments to emphasize the importance of public schooling in shaping immigrant lives; and he summarized evidence from many newspaper accounts using quantitative indices. In this work, the deliberate integration of quantitative and qualitative evidence does not result from a new commitment to “mixed methods research,” but rather from the deep-rooted inclination of historians to use every form of evidence available to construct and test explanations.

The deliberate integration of multiple forms of evidence characterizes other social science disciplines as well. An excellent recent example of the “*ex ante*” blending of quantitative and qualitative evidence in sociology appears in Laub and Sampson (2004), who consider alternative explanations for when and why high-rate offenders desist from criminal activity. Using longitudinal data from ages 15-70 on a sample of delinquent boys growing up in the Boston area, the authors test the theory that social integration in the form of employment and marriage predict dramatic reductions in criminal behavior. Although prior theory predicts that such a process occurs in late adolescence and early adulthood, the broad age-range of their sample enables the authors to show that these dynamics apply across the life course. Laub and Sampson use sophisticated growth curve

analysis with marital status and employment as time-varying predictors of later crime; they supplement evidence from these statistical models with in-depth interviews. The growth curve analysis showed that employment and marriage predict subsequent reductions in crime. The interviews showed how and when a newly married woman's influence becomes sufficiently strong to shape the future behavior of her husband: when men moved in with their wives, they often moved geographically as well, and their routine activities changed, often at the urging of their wives. As a result, the men abandoned social networks engaged in criminal offending, and their own criminal propensities declined accordingly as the geographic and social distance from past co-offenders increased. Laub and Sampson's book is, in fact, the third book to combine in-depth qualitative interviews with longitudinal data analysis to uncover the roots of crime, its persistence, and desistance on this same sample; see also Sampson and Laub (1993) and Glueck and Glueck (1930). Taken together, these three books based on the "Glueck men" constitute a treasure in modern criminology and a 70-year long lesson in *ex ante* mixed methods research.

Nor is the integration of quantitative and qualitative evidence new to classroom research. Rist's (1970) ethnographic account of the formation of teacher expectations during the first days of an inner-city kindergarten is perhaps one of the most widely cited studies in all of educational research. In that work, Rist cites another landmark work on expectancy, Rosenthal and Jacobsen's (1968) widely heralded "Pygmalion" study, an experimental study of the causal effect of teacher expectations on pupil IQ. Rist regarded Rosenthal and Jacobson's study as establishing the existence of the causal effect; he designed his own qualitative study to reveal how teachers formed their expectations

during interactions at the beginning of the year; how these expectations shaped the social organization of the classroom; and how classroom organization shaped learning opportunities. Later, Rosenthal (1972) cited Rist's work in developing a four-factor theory for how teachers communicate their expectations, a theory later tested by Rosenthal and his colleagues but also by Brophy and Good (1971, 1974) using quantitative data in a series of studies that had wide impact on classroom teaching. I am hard-pressed to recall a more influential application of *post hoc* mixed methods than this sequence of investigations.

The Rist study encountered the inevitable (and fair) criticism that life in a single inner-city kindergarten classroom could not be regarded as representative across any larger universe of settings. The Rosenthal and Jacobsen study encountered withering criticism as well (see Elashoff and Snow, 1970), probably more because of the study's visibility and impact than because of its inherent flaws (see Rosenthal and Rubin, 1981; Raudenbush, 1984). Such debates are essentially irresolvable, so entrenched were warring perspectives; nor can a single study be expected to resolve such conflicts. The healthy scientific response was to seek replication.

Several of the first attempts failed to find any effects of the experimentally induced expectations. Yet qualitative interview data published in those studies gave a clue as to *why* the experiments may have failed. The experimental paradigm relied on deception: researchers gave teachers false information in the form of inflated test scores showing which children could be expected to show intellectual growth. If the teachers believed the researchers, they would presumably act on their new positive expectations for these "bloomers," setting in motion the self-fulfilling prophecy whereby positive

teacher expectations encourage positive teacher behavior, in turn motivating a favorable response in the children, whose efforts would rise to meet the high expectations of the teacher, leading to higher test scores. But if the researchers were unsuccessful in deceiving the teachers, the manipulation could not affect the teacher expectations and no treatment would actually be implemented. Thus, no effect would be produced, not because of the failure of the theory underlying expectancy effects, but because of a failure to implement the treatment. In interviews conducted after one of the early replication studies, teachers said that they did not believe the researchers' assertions about which children were going to bloom (Jose and Cody, 1971). The teachers said that they had already known the children well by the time the study began, several months into the school year. According to the interviews, the teachers relied on their own substantial knowledge of the kids rather than on the test data provided by the researchers.

Several of the studies thus differed from the original Pygmalion study in a crucial regard: the timing of "expectancy induction." Pygmalion was launched at the beginning of the school year, before the teachers knew the children well, while several early failed replications confronted the teachers with the inflated test scores mid-way through the year, after the teachers knew the children well. In my synthesis of findings across 18 replications of this experiment (Raudenbush, 1984a; 1984b), I hypothesized that only when the experimental intervention occurred at the beginning of the school year – before the teachers knew the children – would the expectancy effect appear. This turned out to be correct: the timing of expectancy "induction" was a powerful predictor of the experimental effect size. I also used quantitative data from later experiments to test hypotheses generated in (qualitative) criticisms of the earlier studies. In those reviews

(c.f., Elashoff and Snow, 1970; Jensen, 1969) critics hypothesized that expectancy effects would occur only: when the teachers administered the IQ tests; when the children were young and therefore impressionable; and when the tests were group-administered. Fortunately, a large number of later replication studies became available to test these hypotheses. None of these hypotheses were supported, based on the quantitative synthesis of experimental data across studies. The most economical interpretation was that the expectancy effect on IQ is real, but that the deception required to implement the treatment can succeed only if the teachers have not had previous contact with the students.

These case studies illustrate some of the ways in which quantitative evidence and qualitative evidence can productively be exploited to enhance understanding. In Mirel, quantitative evidence showed that Catholic immigrant parents tended most often to send their children to public schools; qualitative evidence from local newspapers characterized experience in those schools. In Laub and Sampson (2004), longitudinal survey data established the statistical association between marriage and subsequent reduction in criminal offending; interviews provided a plausible explanation for this finding, elaborated in detailed case studies. Rosenthal and Jacobsen's experimental study showed an effect of teacher expectancy on pupil IQ; this motivated Rist to use anthropological methods to find out how teachers form their expectations. Rist's findings helped Rosenthal (1974) develop a theory of the transmission of expectations, tested quantitatively by Brophy and Good (1971; 1974) using data from surveys, quasi-experiments, and randomized experiments. Early replications of Rosenthal and Jacobson failed, prompting Jose and Cody to interview teachers, producing hypotheses about when

and why expectancy effects occur, tested by Raudenbush (1984a) using “meta-analysis,” a form of quantitative inquiry that might be termed in this case as a “survey of past experiments.” Qualitative data in the form of criticisms of earlier studies led Raudenbush (1984b) to form and test hypotheses about effects arising in later studies.

I selected these examples, quite frankly, based on personal knowledge rather than based on any systematic search, although at least some of them are important case studies in social science research. There are undoubtedly vastly more examples of this type. All of these examples could be regarded as typical of scientific practice: scientists use a variety of information – clinical and anecdotal, as well as scientific – to generate hypotheses. They choose the most convincing method available to test their hypotheses, and their results generate new questions about when and why an effect appears. They then seize opportunistically on any available and useful approach to answer these questions. Just as Sherlock Holmes would never restrict himself to using a magnifying glass in searching for clues that might solve a crime, scientists, who are essentially professional detectives, would be irrational if they allowed methodological inflexibility to screen out new opportunities for insight.

If mixed methods research is a natural outgrowth of scientific curiosity, why are we now hearing a clarion call for mixed methods in education research in education?

### **Why Are We Talking About Mixed Methods Now?**

Over the past four years, causal questions – questions about the impact of alternative policies and practices – have emerged as priorities in the educational research. These priorities are clearly reflected in the research and evaluation agenda of the US

Department of Education. Questions drive methodological choices, and randomized experiments provide the clearest answers to causal questions arising in social science. It should therefore not be surprising that the Department has developed a strong inclination to fund randomized studies.

At the time of this shift, a reasonably large fraction of the educational research establishment had long been mired in a “paradigm” war over the logic and methods of inquiry in education. Having touched on this above, I will not dwell on the terms of this debate (see Johnson and Onwuegbuzie, 2004, for a review), but there is little doubt that paradigmatic opponents of quantitative research feel isolated by the turn of events at the Department.

A much larger group of educational researchers, who might be characterized as “pluralists,” have historically rejected the notion that there are viable alternative paradigms for educational research and have reasoned, instead, that a variety of questions are of interest in educational research, that methodological choices should be tailored to questions, and, therefore, that educational researchers should be trained to understand and respect alternative approaches and methods of inquiry. This pluralist perspective is vividly articulated by Shulman (1988) who built on Cronbach and Suppes’ (1969) notion of “disciplined inquiry” as a unifying concept capable of embracing a broad range of questions and methodological orientations:

“Disciplined inquiry has a quality that distinguishes it from other sources of opinion and belief. The disciplined inquiry is conducted and reported in such a way that the argument can be painstakingly examined. The report does not depend for its appeal on the eloquence of the writer or on any surface plausibility (Cronbach and Suppes (p 15), quoted in Shulman (1988)).”

“What is important about disciplined inquiry,” Shulman writes, is that its data, arguments, and reasoning be capable of withstanding scrutiny by another member of the

scientific community.” Why, then did Cronbach and Suppes, and later Shulman, substitute the term “disciplined inquiry” for “science?” My reading of these works suggests that the canons of disciplined inquiry are similar to the canons of science, but the aim was create a foundation broad enough to encompass historical research, ethnography, and interview studies – forms of inquiry that at times straddle the boundary between social science and the humanities. The aim was to foster a broad, publicly accountable community of scholars committed to transparent assumptions, reasoned criticism, and, whenever possible, publicly available data. Within this broad community, it would be possible to exchange new findings and understandings and to raise new questions, creating a more productive intellectual life than that envisioned by the paradigm warriors.

One way to interpret the call for mixed methods as embodied in this Forum is as a plea from this pluralist perspective in the face of current efforts, at the Department of Education but also more broadly among policy makers and research funders, to emphasize causal questions and randomized experiments. This appears to be the spirit of Schoenfeld (2004), when he writes

“Even today there are large numbers of funders and researchers who do not understand that there are many ways of conducting rigorous, high quality research in education, and that appropriately used statistical methods are only one category of such work (page 7).”

But an argument for multiple methods, whether used in separate studies or integrated within the same study, can be evaluated only in the context of clearly defined research questions. As everyone seems to say, questions should drive methods; and in the context of constrained resources, only some questions can be pursued. Thus, it is essential

to articulate a compelling research agenda before evaluating the role that multiple methods might play in reinforcing the scope and credibility of any research effort.

In the next section, I argue that the question before us now is not whether and how to employ mixed methods in educational research generally; rather the question is whether and how to employ these methods in service of a newly dominant research agenda that aggressively seeks to evaluate claims about the causal effects of interventions aimed to improve teaching and learning in the nation's classrooms.

## **II. Setting Priorities**

Among policy makers, public and private research funders, and applied educational researchers themselves, there is currently an unmistakable and perhaps overarching interest in identifying interventions that show strong promise, based on convincing evidence, to improve teaching and learning in US classrooms. These interventions might include new curricula, new technologies, new instructional methods, new forms of teacher preparation and in-service training, and new ways of organizing schools to support effective practice. The sources of this interest are not hard to identify.

1. A wealth of evidence reveals large and persistent gaps in literacy between high- and low-income children and between white children, on the one hand, and African American and Hispanic children on the other. The US government's primary intervention into schools over the past 40 years, better known as Title I, has aimed specifically to raise the achievement of low-income children, yet the quality of available evidence on how best to achieve that goal is remarkably thin. In this context, it is not surprising that the

evaluation agenda at the US Department of Education emphasizes the development of new knowledge about how best to intervene to solve this problem.

2. Prior research shows unmistakably that US students of all social backgrounds score disappointingly low on international assessments of mathematical and scientific knowledge. Yet despite large investments in inventing new curricula, new technologies, and new approaches to instruction and teacher training, reliable knowledge on how to improve learning in math and science in US classrooms remains weak. Not surprisingly, the National Science Foundation has launched new efforts to develop evidence about the effectiveness of innovative classroom interventions “at scale,” meaning in regular classrooms on a large enough scale to make a practical difference.

Of course policy makers are not well-positioned to intervene in classrooms to improve instruction. But every major policy initiative relies on the assumption that practitioners will understand how to intervene if supplied with sufficient resources and incentives. Consider the three major policy initiatives currently under discussion: providing more resources, increasing accountability, and transforming school governance.

### **Increasing Resources**

One option to improve learning is to make more resources available. Presumably, spending more money per child, increasing teacher pay and qualifications, building better facilities, investing in technology, and reducing class size will boost student learning. Not surprisingly, evaluating the effects of investing resources has been a major pre-occupation of educational research at least since the “Coleman report” (Coleman et al., 1966). Most reviews of the evidence, however, are not encouraging (Hanushek, 1989).

Certainly this body of work has revealed evidence of some effects (Greenwald, Hedges, and Laine, 1996) and particularly in regard to class size reduction (Finn and Achilles, 1990; Nye, Hedges, and Konstantopoulos, 2000; Krueger and Whitmore, 2001). It is hard to assert, however, that this work has had more than a marginal impact on the quality of classroom learning in the US.

Cohen, Raudenbush, and Ball (2003) reason that such resources are used in so many ways with such varied effects that it is difficult to predict the outcomes of investing in them. These authors argue for a shift in how we conceive the study of resources. Rather than conceiving resources as the causal agent and achievement as the outcome, they advocate a research agenda in which well-defined “instructional regimes” are the causal agents. Having discovered an effective regime, the next logical task is to assess how constrained resources might affect the impact of the regime. In particular, will an instructional regime found successful in a small pioneering study continue to demonstrate success if, when taken to scale, classes are a bit larger, teachers are somewhat less well-trained, or facilities are somewhat less than optimal than in the original study? Such an approach has parallels in the history of medical research: clinical trials focus on the efficacy of new clinical practices, while health services researchers study ways to make effective clinical practices broadly available.

The key point is that a policy of investing in resources to boost achievement assumes that practitioners will know how to use those resources in instruction. This requires not only that teachers know how to use the new resources in their instruction; it also requires that district and school leaders will understand how to use the new resources in coordinating instruction across the grades and across the schools their students are

likely to attend. Given the current weakness in knowledge about how best to organize instruction, it seems hardly surprising that simply investing in new resources would have at best, marginal effects on student outcomes.

### **Increasing Accountability**

The second major policy tool for improving learning has been to increase accountability. A considerable emphasis in federal and state policy over the past 20 years, culminating in No Child Left Behind (“NCLB”) legislation, is based on the following theory of action: hold educators accountable for student outcomes based on state assessments, but give these educators wide discretion in devising the means to produce those outcomes. To be successful, the approach must motivate educators to pursue goals embodied in the assessments; and those educators must find effective means to achieve those goals. It seems clear that school people are now sufficiently motivated. But will educators have adequate knowledge to select interventions that will improve learning? Perhaps, but the amount and quality of evidence those practitioners can draw on is, by all accounts, weak.

Apparently, then, more knowledge about how to improve instructional practice is the critical missing ingredient in the success of the accountability reform. While many have argued for injecting more resources under NCLB, we have already considered the problem with that argument: resources, by themselves, do not improve teaching and learning. Knowledge about how to use those resources in instruction is key, yet that knowledge is woefully lacking.

## **Reforming Governance**

The third major policy initiative aimed at improving student outcomes is a transformation in school governance by means of school choice plans, whether enacted through privatization, charter schools, or some other mechanism. The theory of action here is that competitive pressures will produce incentives for school improvement in order to attract customers (parents); and that freeing educators from the bureaucratic constraints of the conventional local education authorities will give educators the flexibility to modify practice to produce high-quality instruction capable of attracting these customers. Assumptions are that customers will know quality when they see it, and that educators, free of bureaucratic constraints, will know how to create quality. Once again, the knowledge gap looms large, and the key task for educational research is to produce reliable evidence about instructional interventions, or “instructional regimes” (Cohen, Raudenbush, and Ball, 2003). Giving educators the flexibility they need to adopt effective practices is an admirable aim, but knowledge of which practices are effective is essential if this kind of reform is to affect teaching and learning in powerful ways.

## **Conclusion**

In sum, policy makers could not directly intervene to improve instruction, even if they knew how to do improve it. What they can do is supply resources and incentives, with accountability and governance reforms exemplifying two potentially linked strategies for shaping incentives. But effective instruction is not likely to flow automatically from exerting these policy levers any more than giving doctors resources and incentives to save lives will produce optimal medical practice. A knowledge gap needs to be addressed, so that educators can act on incentives and use resources in ways

that will supply students with coherent and effective instruction. *It follows that identifying, testing, and warranting the effectiveness of strategies for instruction is currently the central task of applied research in education.*

This is not to say that more basic research – in the history, sociology, politics and anthropology of education, or in cognitive and neuro-science – are not terrifically important and, ultimately, potentially useful for policy and practice. It is to argue, however, that within the domain of applied educational research, that is, research linked closely to current problems of policy and practice, priorities must be set; and that all roads seem to lead to instructional improvement as the central priority at this time.

The next task of this essay is to identify the key questions that emanate from such an agenda and to define broadly the methodological priorities that follow from these questions with an eye to assessing the role of mixed-methods research.

### **III. Questions and Methods**

I have argued that instructional improvement is central to the nation's applied educational research agenda. We must identify instructional regimes – coherent approaches to organizing instruction in the domain of specific subject areas – that can be implemented on a broad scale and that can be relied upon to produce good impacts over a comparatively broad range of conditions for well-defined target populations. Such a research agenda must produce strong warrants about the causal effects of implementing these regimes. Let us now consider the kinds of studies needed to support this agenda; as a corollary, to let us identify the methodological approaches entailed in this effort. Within

this scenario, what is the contribution of mixed methods? Specifically, how might various quantitative and qualitative data collection efforts combine to accomplish this goal?<sup>1</sup>

The problem of identifying effective instructional regimes is, to use current jargon, a question about “what works.” Such questions are essentially causal questions, and social scientists generally regard well-planned experiments as the best way to discover the causal effects of alternative innovations. Not surprisingly, therefore, I advocate systematic experimentation as central to the research agenda. Experiments, while necessary, are, however, far from sufficient to support the learning required for effective instructional innovation. Other kinds of research are needed to precisely define educational aims, to identify target populations for intervention, to identify most promising practices, and to clarify challenges and opportunities for effective implementation of those practices. The challenge we face in promoting a seemingly diverse research agenda is to get clear on how these efforts can be integrated to support the broad goal of discovering and warranting best practice.

### **Why Experiments *Are* the Gold Standard for Causal Inference**

Statistical science (c.f., Rubin, 1978; Holland, 1986), paralleling similar developments in economics (see Heckman, 2004) has come to define a causal effect for a given child as the difference between two potential outcomes: the outcome the child

---

<sup>1</sup>Some readers will disagree, of course, about my reading of the research priorities. Yet I would hope that most would agree that to answer questions about the optimal mix of methodological approaches requires *some frame*, some set of orienting problems and questions that require a methodological response. Otherwise, the discussion will be empty or worse, misleading, as question-free methodological discussions will typically be. It would be interesting to see the implications of alternative research agendas for methodological choices and, in particular, for assessing the role of mixed methods.

would display if one course of action were followed (e.g., if the child were to experience a novel approach to promoting reading comprehension) minus the outcome that same child would display if, instead, some other course of action were followed (e.g., if the child were to receive the current reading program). Such a causal effect can never be observed, since it is impossible for a child to receive both interventions at the same time. However, it is possible to estimate the *average causal effect* for a population of children, or for some sub-population, *under assumptions*. A key assumption is that each child's assignment to one intervention or the other does not depend on that child's potential outcomes. Statisticians refer to this as the assumption of *ignorable treatment assignment*. This means, in part, that the children assigned to receive a new experimental intervention would have displayed the same average outcome as control-group children had those experimental-group children instead been assigned to the control group. A second aspect of this assumption, often ignored in discussions of causation, is that those children who stand to benefit most from the new intervention are neither more likely nor less likely to receive it.

The random assignment of children, classrooms, or schools to alternative interventions insures the validity of the assumption of ignorable treatment assignment. For example, if the flip of a coin determines the assignment of a school to an experimental or control group, every child has a probability of 1/2 of receiving the experimental intervention. Thus, the child's potential outcomes cannot predict treatment group assignment.

Moreover, in a randomized experiment, conventional significance tests and confidence intervals quantify the researcher's uncertainty about the existence and

magnitude of the causal effect. Put more simply, it is true that, by chance, differences will exist among randomly formed groups, and these differences may, in fact, be quite large in small samples. But such chance differences are fully accounted for by well-known and comparatively simple methods of statistical inference.

While school-based randomized experiments have been comparatively frequent in public health, including, for example, research on interventions aimed to reduce violence or substance use, such studies have, until quite recently, been comparatively rare in evaluations of interventions designed to improve teaching and learning (Cook, 2000). This means that evaluators of educational innovations have had to rely on non-experimental methods of attempting to satisfy the assumption of ignorable treatment assignment. These include, prominently, quasi-experimental designs (Shadish, Cook, and Campbell, 2001) combined with statistical control for potentially confounding variables (c.f., Rosenbaum and Rubin, 1983). A confounding variable is a characteristic of a student, classroom, or school that predicts treatment group assignment and also predicts potential outcomes. Failure to control for such variables, known as “confounders” for short, has plagued many past evaluations. For example, the first evaluations of Head Start in the 1970s found no significant mean difference in cognitive outcomes between those who children who did and those children who did not experience Head Start. This led readers to conclude that Head Start was ineffective. However, critics pointed out that the children receiving Head Start were significantly more disadvantaged than the comparison group on family education and income. Thus, it was plausible to predict that Head Start children would have done worse, on average, than the comparison group even in the absence of the program. If so, the failure to control for confounders would bias the

evaluation against Head Start. One might even speculate that the failure to find a difference between the two groups indicated a *positive* effect of Head Start, though such reasoning remains speculative in the absence of a more rigorous strategy for eliminating confounding.

In light of such painful experiences, it is not surprising that educational evaluators using non-experimental methods have become ever more sophisticated in their attempts to identify and control for confounders. The challenge they face is a tough one: no matter how many potential confounders they identify and control, the burden of proof is always on the evaluator to argue that no important confounders have been omitted. Perhaps the chief strategy in studies of interventions aimed to increase achievement has been to insure that students are administered a reliable pre-measure of the same achievement variable to be used as the outcome. Presumably, much of the association between a potential confounder (e.g., an aspect of home environment) and the outcome is removed once one has controlled for a reliable pre-test of achievement. Substantial experience supports this basic idea.

Recall, however, that the assumption of ignorable treatment assignment has two parts. One part is that more able students are no more or less likely than less able students to receive the new intervention. Put another way, the two groups would have had the same average achievement if both groups had received the “control” treatment. It seems reasonable that adjustment for a good measure of prior achievement would “soak up” much of the bias thus conceived.

However, the second part of the assumption of ignorable treatment assignment is that one’s potential to benefit from the treatment is unrelated to treatment group

assignment. This means that researchers who do not use randomized assignment must identify and control for pre-treatment characteristics of children, classrooms, and schools that pre-dispose children to *benefit from the treatment*. This requirement poses a major challenge to valid inference, especially in cases where agents such as administrators, teachers, or parents, or even the children themselves select which treatment the children will experience. Those agents may have information on the potential benefits of selecting the treatment, information unavailable to the researcher and thus incapable of incorporation into the quasi-experimental design or the statistical analysis. In this case, pre-treatment matching or statistical control for measured confounders would not be sufficient to remove bias. Random assignment solves this problem.

Critics have argued that randomized studies may be unethical or difficult to pull off in educational settings where agents such as principals, teachers, parents, and even students may have fairly substantial autonomy. Recent experience suggests that, as a general proposition, this argument is unfounded. A thoughtful design phase that incorporates the needs and concerns of local actors can often produce a successful randomized experiment. Experience shows that teachers and school leaders will participate in group-randomized studies when they are convinced something important can be learned about how to improve teaching and learning and when the study does not threaten their basic interests.

The recent experience of Robert Slavin and colleagues (2004) is instructive. Initial attempts to recruit schools to participate in a randomized study of “Success for All” were fruitless. School leaders resisted participating in a study that could result in their school landing in the control group. Their concern was sensible in light of current

pressure on schools to improve under NCLB. So Slavin re-designed the study. In one half of the schools, Success for All would be implemented in kindergarten during first year, kindergarten and first grade during the second year, and kindergarten through grade 2 in year 3. In a second random half of the schools, Success for All would be rolled out in grades 3-5 during the first three years. Ultimately, all schools would receive the program at all grade levels. But during the first three years, each school receiving Success for All in grades 1-3 would supply control group data from its grade 3-5 students; similarly schools receiving the program in grades 3-5 would produce control group data from its k-2 students. Using this strategy, in which all participating schools stood to benefit from a new intervention while also contributing to new knowledge, Slavin and colleagues were able to recruit 40 schools for the study, a sufficient number to insure adequate statistical power.

Based on these and other experiences, randomized experimentation has emerged not only as the logically optimal approach to valid causal inference but also as ethically and practically viable under a reasonably broad range of circumstances. Clearly, randomized experiments ought to play a central role in a research agenda designed to discover and disseminate effective new interventions for instructional improvement.

Nevertheless, such experiments cannot be regarded as sufficient to insure the success of this research agenda. I now consider the complementary studies and appropriate research methods needed to insure success in this endeavor.

## Why Experiments Are Not Sufficient for Our Agenda

Our proposed research is constructed to study these interventions in order to learn what works so that educational policy and practice can be based on the best available evidence regarding promising new innovations. Because randomized experiments are the best way, in principle, to discern the causal effects of such interventions, it may therefore appear that a well-planned sequence of randomized experiments would suffice to achieve our aims. While this simple reasoning has a surface appeal, I reject it. Well-designed randomized experiments are, I believe, necessary but not sufficient for our purpose.

The argument is straightforward and perhaps obvious. Innovations in curriculum, instructional technology, and teacher professional development are *interventions* designed to improve *outcomes* for particular kinds of *kids* in specific *settings*. For our research agenda to succeed, we need considerable precision in defining the outcomes we want to pursue. We need to identify interventions that hold greatest promise in achieving those outcomes. We need good data to decide which kids to target in which settings because kids with particular needs are of greatest importance in particular instructional situations. For example, third graders who have failed to respond well to good reading instruction will likely require a different remedial intervention than will third graders who have never experienced adequate reading instruction.

The randomized experiment becomes a powerful tool for warranting causal effects after a rather protracted process has identified the most promising interventions to change the most important outcomes for target kids in settings of interest. This process involves a series of well-designed descriptive and correlational studies using a variety of

methods, quantitative and qualitative, without which the program of randomized experimentation is doomed to failure.

One might ask: why not use a randomized experiment to test the effects of every potentially interesting intervention on every possibly relevant outcome for every important target population? With an infinite research budget and limitless prior knowledge about how to implement a given intervention in the turbulent setting of classrooms and schools, this might be a good idea. However, in the world as it exists, experiments are quite expensive relative to available funds for research, and we actually know little about how to implement a new intervention until we have tried doing so, at least in small scale settings. It therefore makes sense to insure that an intervention is capable of successful implementation on a broad scale before submitting that intervention to a randomized trial of effectiveness. Testing good ideas that are poorly implemented does not tell us “what works.” Moreover, a series of large-scale experiments testing poorly conceptualized programs represents a serious waste of resources. For these reasons and more, a multifaceted research agenda is essential to support systematic experimentation.

Below I sketch some of research that must accompany and support a well-planned series of experiments. I emphasize not only the diversity of research approaches but also the crucial question of how these must be integrated if the entire program of research is to succeed.

### **Defining Relevant Outcomes**

Large-scale assessments provide detailed pictures of what American youngsters know and can do in core subject areas. NAEP, TIMMS, and state assessments come

quickly to mind, but many other studies assess aspects of children's conceptual understanding, procedural knowledge, and content knowledge in mathematics and science as well as their phonemic awareness, vocabulary, reading fluency, and comprehension. These studies identify gaps in student proficiency that ought to motivate critical examination of practice and spur innovative program design. Such assessments are essentially surveys built upon accumulated knowledge from cognitive science, expert judgment, and psychometrics. Without them, policy makers and researchers would not be clear on which outcomes for which kids are in greatest need of improvement.

More broadly, innovative thinking often entails new goals for student learning. These new goals, by definition, are not operationalized in off-the-shelf tests. If these new goals are to be pursued and assessed, they must be made precise, laying a basis for new test construction. Test construction is a complex business, entailing new frameworks, new tasks or items, new ways of summarizing evidence about student proficiency, and field tests of reliability and validity. The invention of new goals, the construction of new tests, and their validation is itself an ambitious program of research that requires a mix of qualitative and quantitative inquiry as we study how students respond to new tasks and make meaning of potential items. This process of generating valid new assessments of student learning requires considerable new psychometric investigation as well.

Indeed, one might argue that a failure to attend systematically to this process of creating good outcome measures is the achilles heel of evaluation research on instructional innovation. If this process is ignored, trivialized, or mismanaged, we'll be measuring the wrong outcome with high reliability, the right outcome with low reliability, or, in the worst case, we won't know what we are measuring. If we don't know

what we are measuring, the causal question (Does the new intervention improve achievement?) is meaningless. If we measure the right outcome unreliably, we will likely find a new program ineffective even if it is effective. If we measure the wrong outcome reliably, we may find that the intervention "works" but we'll never know whether it works to achieve *our* goals.

### **Identifying Promising Interventions**

As mentioned earlier, there are many more potentially interesting programs than there are resources needed to evaluate them with randomized experiments. Thus expert knowledge, attempts to implement novel programs on small scales, and preliminary (non-randomized) assessments constitute an array of strategies to discard some interventions and to refine others before summative tests of effectiveness ought to be tried. Detailed descriptions of expert practice often supply key new ideas for how to intervene. Small-scale implementation studies or even careful small-scale randomized studies can provide preliminary evidence about whether a new approach can, under ideal conditions, produce an effect for a sample that probably isn't representative. Secondary analysis of large-scale data can provide important evidence of promising practice. The synthesis of research from a variety of methods conducted at different scales ought to be a pre-requisite for the construction of a large-scale randomized field trial.

An example of secondary analysis informing intervention comes from TIMMS. Schmidt and McKnight found that, compared to mathematics instruction in the US, which tends to cover many topics over comparatively short periods of time, instruction in several other nations tends to be highly focused on mastery of a few topics over an

extended period of time. The authors characterized the instruction in such countries as more focused and coherent than the instruction in the US. This evidence, drawn from a large scale survey of 50 countries was supported by in-depth analysis of videotapes of representative instructional scenarios in the several countries. It turns out that the children in nations pursuing more focused and coherent instruction do substantially better on the mathematics assessments than do US children. Further preliminary research might involve secondary analysis of US data bases combined with case studies to see whether some US teachers pursue focused and coherent instruction (as defined by Schmidt and McKnight using TIMMS) and whether students in those classes fare well. Together, these findings, culled from a variety of descriptive methodologies quantitative and qualitative, in principle lay the basis for the invention of a new innovation that could be constructed and tested on a small scale. The next logical step would be evaluation by means of a randomized trial, which would supply the strongest possible evidence about causal effects in a US context.

### **Targeting populations of Interest**

Whose outcomes are we aiming to improve? Many researchers are interested in overcoming achievement gaps, and in fact, doing so is an official goal of Title I, the largest federal program in K-12 education. But how do we know that such gaps exist? How do we know whether those gaps are already diminishing over time? Once again, a variety of research has been essential to find out which kids are faring well and not so well, whether gaps are increasing or shrinking, and whether available data contains clues regarding the types of settings, organizational approaches, and strategies for instruction that might most plausibly help overcome those gaps.

We need to know whether the most disadvantaged kids lack good teachers or other resources, in which case equalizing resources might be a promising strategy. Alternatively, such kids may thrive in instructional environments that are not effective for other kids. In this case, equalizing resources may not be the answer. Instead, the answer may involve tailoring instruction to the specific needs of these children. Once again, a variety of research strategies, ranging from large-scale surveys to small-scale qualitative observation and interviewing, is important to answer these questions, and answering these questions is potentially important for the design of field trials of innovations.

Other targets for intervention might be second-language learners, children with disabilities, girls, or those demonstrating early potential to become top mathematicians. In each case, research evidence is essential in designing relevant options for policy and practice, options that can in many cases be tested by means of experimentation.

### **Putting the Pieces Together**

In sum, a well-planned strategy of experimentation is optimal in generating solid evidence about the likely impact of new innovations in school organization, curriculum, instructional technology, and professional development. The Department of Education, the National Science Foundation, the National Institute of Child Health and Human Development, along with a number of private foundations systematically support the generation of promising innovations, and it makes great sense for these agencies to support an ambitious program of evaluation research to insure that the nation learns from attempts to improve teaching and learning. I have summarized key arguments in favor of

random assignment of schools, classrooms, or students to alternative instructional programs and found the case for randomized studies compelling.

At the same time, I have argued that experimentation, while necessary, is far from sufficient to achieve the goal of learning about "what works." Research using a variety of methods is essential:

a) to define the student outcomes we seek to change and to build and validate assessments of those outcomes;

b) to support novel thinking about how best to intervene, to support preliminary studies of those interventions, and to enable educators to test the feasibility of implementing those interventions in ordinary school settings;

c) to clarify the subsets of kids who are in greatest need of intervention or who are most likely to benefit from new ideas about teaching and learning.

A final goal is to study why an intervention works, why it works for some kids and not others, or why it fails. A variety of methodological strategies, including qualitative and quantitative studies of implementation, interviews of teachers and children, and observations of practice can produce plausible explanations, new hypotheses, and ideas for refining interventions. Descriptions of practice in "settings of origin" (that is, settings in which a new intervention is initially found effective) can be compared to descriptions of practice when the intervention is implemented on a broader scale. Such comparisons can reveal the extent to which practice has shifted under the impact of exigencies not present in the original setting, laying a basis for understanding the deterioration of an effect as it is taken to scale and suggesting ways to strengthen training and administrative support.

## **Implications for the Support of Mixed Methods Research**

The success of an ambitious program of innovation and experimentation appears to depend on a fairly complex set of inter-related research activities. These are required to refine aims and develop outcome variables; to identify or invent promising innovations; to study the feasibility of implementation; to test causal impacts at a larger scale; to explore why the intervention works and for whom; and to investigate cost effectiveness by probing how resource constraints bear on the impact of the intervention. My purpose is not to spell out in any detail how researchers might self-consciously employ a diversity of methodological strategies to achieve these aims. Rather, I have attempted to make the case for this effort, in support of a research agenda in which randomized experimentation plays a central role.

The effort thus sketched assumes a fairly cohesive scholarly community in which information and criticism flow rapidly across disciplines and methodological specializations. It requires research training that enables newly minted educational researchers to read and critically evaluate research findings from a wide range of methods while being expert in a specific methodological orientation. The effort requires research managers and funders who can keep in mind the broad aim – to improve achievement by improving teaching and learning in classrooms – while understanding the complementary efforts that must contribute to its achievement.

Finally, the effort thus sketched requires adequate funding. By now the small fraction of all funding for education that supports educational research is well known. It

is hard to imagine how an ambitious agenda of randomized trials supported by a multi-disciplinary and multi-methodological effort can succeed at present funding levels. The weak yield of applied research has undermined the case for generous funding of educational research. On a more hopeful note, successful efforts to produce sound evidence about how to intervene should, in principle, generate wider support for educational research, thereby increasing capacity to mount an increasingly ambitious and effective research effort.

### References

Brophy, J.E. and Good, T.L. (1971). *Looking in Classrooms*. New York: Harper and Rowe.

Brophy, J.E., and Good, T.L. (1974). *Teacher-student Relationships: Causes and Consequences*. New York: Holt, Rinehart, and Winston.

Cohen, D.K., Raudenbush, S.W., & Ball, D.L. (2003). Resources, instruction, and research. *Educational Evaluation and Policy Analysis*,

Coleman, J.S., et al. (1966). *Equality of Educational Opportunity*. Washington: US Department of Health, Education, and Welfare, Office of Education.

Cook, T.D. (2001). Considering the major arguments against random assignment: An analysis of the intellectual culture surrounding evaluation in American schools of education. In R. Boruch and F. Mosteller (Eds.) *Education, Evaluation, and Randomized Trials*. Brookings.

Cronbach, L.J., and Suppes, P. (1969). *Research for Tomorrow's Schools: Disciplined Inquiries for Education*. New York: MacMillan.

Elashoff, J. and Snow, R. (1971). *Pygmalion Reconsidered*. Worthington, Ohio: Charles A. Jones.

Engels, F. (1844). *The Condition of the Working Class in Britain*. (English edition published in 1887 in London.)

Finn, J.D., and Achilles, C.M. (1990). Answers about questions about class size: A statewide experiment. *American Educational Research Journal*, 27, 557-577.

- Greenwald, R. Hedges, L.V., & Laine, R.D. (1996). The effect of school resources on student achievement. *Review of Educational Research*, 66, 361-396.
- Glueck, S. and Glueck, E. (1930). *500 Criminal Careers*. New York: Alfred A. Knopf.
- Hanushek, E. (1989). The impact of differential expenditures on school performance. *Educational Researcher*, 18(4), 45-51.
- Heckman, J. (2004). *The Scientific Model of Causality*. Occasional paper, University of Chicago Department of Economics.
- Jensen, A. (1960). How much can we boost IQ and scholastic achievement? *Harvard Educational Review*, 39: 1-123.
- Jose, J. and Cody, J. (1971). Teacher-pupil interaction as it relates to attempted changes in teacher expectancy of academic achievement. *American Educational Research Journal*, 8, 39-49.
- Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396), 945-960.
- Johnson, R.B., and Onwuegbuzie, A.J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher*, 33(7), 14-26.
- Krueger, A., and Whitmore, D. (2001). The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from Project Star. *Economic Journal*, 111,1-28.
- Laub, J. and Sampson, R.J. (2004). *Shared Beginnings, Divergent Lives: Delinquent Boys to Age 70*.
- Lincoln, Y.S., and Guba, E.G. (1985). *Naturalistic Inquiry*. Beverly Hills: Sage.
- Mirel, J. (2004). *Public Schools and the Americanization of European Immigrants* (2004). Colloquium, University of Michigan School of Education, based on a forthcoming book to be published by Harvard University Press.
- Nye, B. Hedges, L.V., & Konstantopoulos (2000). The effects of small classes on achievement: The results of the Tennessee class size experiment. *American Educational Research Journal*, 37, 123-151.
- Phillips, D.C. (1983). After the wake: post-positivistic educational thought. *Educational Researcher*, 12(5), 4-12.

- Raudenbush, S.W. (1984b). Utilizing controversy as a source of hypotheses for meta-analysis: The case of teacher expectancy effects on pupil IQ. In R.J.Light (Ed.) *Evaluation Review Studies Annual*, 303-326.
- Raudenbush, S.W. (1984a). Magnitude of teacher expectancy effects as a function of the credibility of expectancy induction: A synthesis of findings from 18 experiments.
- Rist, R. (1970). Student social class and teacher expectancies: the self-fulfilling prophecy in ghetto education. *Harvard Educational Review*, 40, 411-451.
- Rosenthal, R. (1974). On the social psychology of the self-fulfilling prophecy: further evidence for Pygmalion effects and their mediating mechanisms. New York: MSS Modular Publications.
- Rosenthal, R., and Rubin, D.B. (1971). Pygmalion reaffirmed. In J.D. Elashoff and R.E. Snow (Eds.) *Pygmalion Reconsidered*. Worthington, Ohio: Charles A. Jones.
- Rosenthal, R. and Jacobson, L. *Pygmalion in the Classroom*. New York: Holt, Rinehart, and Winston.
- Rosenbaum, P., and Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 17, 41-55.
- Rubin, D.B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, 6, 34-58.
- Sampson, R.J., and Laub, J. (1993). *Crime in the Making: Pathways and Turning Points Through Life*. Cambridge: Harvard University Press.
- Schoenfeld, A.H. (2004). *Instructional Research and Practice*. A draft essay for the MacArthur Network on Teaching and Learning.
- Shadish, W.R., Campbell, D.T., & Cook, T. D. (2001) *Experimental and Quasi-experimental Designs for Research*. Houghton Mifflin.
- Shulman, L. (1988). Disciplines of inquiry: An Overview. In R. Jaeger (Ed.) *Complementary Methods for Research in Education*, Washington, DC: the American Educational Research Association.
- Slavin, R. (2004) and Borman, G. (2004). Preliminary Results of the Experimental Evaluation of Success for All. Paper presented at the national conference on the Design and Analysis of Group-Randomized Experiments, Ann Arbor: University of Michigan, July.