

Educational Researcher

<http://er.aera.net>

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


Stephen W. Raudenbush
EDUCATIONAL RESEARCHER 2005; 34; 25
DOI: 10.3102/0013189X034005025

The online version of this article can be found at:
<http://edr.sagepub.com/cgi/content/abstract/34/5/25>

Published on behalf of



By

 SAGE Publications

<http://www.sagepublications.com>

Additional services and information for *Educational Researcher* can be found at:

Email Alerts: <http://er.aera.net/cgi/alerts>

Subscriptions: <http://er.aera.net/subscriptions>

Reprints: <http://www.aera.net/reprints>

Permissions: <http://www.aera.net/permissions>

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

by Stephen W. Raudenbush

Education research is an interdisciplinary effort long characterized by methodological diversity. Why, then, do we hear an urgent call for mixed methods now? Apparently, a recent shift in the applied research agenda has fostered concern that methodological pluralism is at risk. In this article, the author argues that (a) a focus on evaluating the effects of instructional interventions is entirely appropriate given current policy dilemmas; (b) randomized experiments are the gold standard for assessing these effects; but (c) the success of the effort depends on a well-integrated, methodologically diverse research effort. He sketches how diverse methods might be combined and how a healthy scientific community might collaborate to generate adequate funding to support this vital enterprise.

Education may be regarded broadly as the process by which people learn the language, ideas, meanings, expectations, and knowledge that 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

The Research News and Comment section publishes commentary and analyses on trends, policies, utilization, and controversies in educational research. Like the articles and reviews in the Features and Book Review sections of *ER*, this material does not necessarily reflect the views of AERA nor is it endorsed by the organization.

long used the enormous array of methodological approaches characteristic of those disciplines as they study education. Against this background of disciplinary and methodological diversity, an outside observer might be puzzled to hear the current call for the use of “mixed” or “multiple” methods in education research (c.f. National Academy of Science, Center for Education, 2005; Johnson & Onwuegbuzie, 2004). What does this call mean and why do we hear it now?

Over the past 4 years, causal questions—questions about the impact of alternative policies and practices—have emerged as priorities in education research. These priorities are clearly reflected in the research and evaluation agenda of the U.S. 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 U.S. Department of Education has developed a strong inclination to fund randomized studies.

One way to understand the current call for mixed methods in education research is as an expression of concern about this trend among a broad range of education researchers who use diverse methods in their work. Members of this group tend to believe that education research will be strong when inferences based on one form of inquiry are checked or “triangulated” against inferences based on an alternative form of inquiry characterized by complementary strengths. Johnson and Onwuegbuzie (2004) clearly articulate this viewpoint. The concern is that a strong emphasis in research funding on causal questions and randomized experiments may undermine the virtues of combining mixed methods of inquiry.

But an argument for multiple methods 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 following pages, I argue that the question before us now is not whether to employ mixed methods in education research generally; rather, the question is how to employ them in the service of a newly dominant research agenda that seeks to evaluate claims about the causal effects of interventions aimed to improve teaching and learning in the nation’s classrooms. I support the argument that causal effects of instructional interventions belong at the heart of the current policy research agenda in education and that randomized experiments are the best way to warrant these effects. However, I also claim that such experiments are insufficient to achieve the aims of this agenda, and I sketch a strategy for combining diverse methodological approaches to achieve these aims.

Setting Priorities

Among policymakers, public and private research funding agencies, and applied education researchers themselves, there is currently an overarching interest in identifying interventions that show strong promise, based on convincing evidence, to improve teaching and learning in U.S. classrooms. Such interventions might include new curricula, new technologies, new instructional methods, new forms of teacher preparation and inservice training, and new ways of organizing schools to support effective practice. The sources of this interest are not hard to identify.

First, 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 U.S. government's primary intervention in 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 (Borman, Hewes, Overman, & Brown, 2003). In this context, it is not surprising that the evaluation agenda at the U.S. Department of Education emphasizes the development of new knowledge about how best to intervene to solve this problem.

Second, prior research shows unmistakably that U.S. 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 U.S. 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 policymakers are not well positioned to intervene in classrooms. Instead, they attempt to influence teaching and learning indirectly by providing resources, increasing accountability, and transforming school governance. Yet the success of such policy initiatives depends critically on knowledge about how to improve instruction. Let us briefly consider each of these policy options.

Providing 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 preoccupation of education 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, & Laine, 1996) and particularly in regard to class size reduction (Finn & Achilles, 1990; Krueger & Whitmore, 2001; Nye, Hedges, & Konstantopoulos, 2000). It is hard to assert, however, that this work has had more than a marginal impact on the quality of classroom learning in the United States.

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. A regime is a coherent plan for introducing content, assessing student progress, and interacting with students to achieve clearly defined goals. Professional development is typically required to ensure adequate subject matter and pedagogical knowledge to enact a "regime" (see Borko, 2004, for further development of this idea).

Having discovered an effective regime, the next logical task is to assess how constrained resources might affect its impact. In particular, will an instructional regime that is 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 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, whereas 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 teachers will know how to use those resources in instruction. It also assumes that district and school leaders will understand how to use the new resources in coordinating instruction across grades and across the schools that their students are likely to attend. Given the current weakness in knowledge about how best to organize, coordinate, and enact effective instruction, it is 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 the No Child Left Behind (NCLB) legislation, is based on the following approach: Hold educators accountable for student outcomes based on state assessments, but give the educators wide discretion in devising the means to produce those outcomes. To be successful, this approach must motivate educators to pursue goals embodied in the assessments; and the educators must find effective means to achieve the goals. It seems clear that those who are involved in schooling are now sufficiently motivated. But will they have adequate knowledge to select interventions that will improve learning? Perhaps, although the amount and quality of evidence that they 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. Although many have argued for injecting more resources under NCLB, we hold that resources, by themselves, do not improve teaching and learning. Knowledge about how to use resources in instruction is key, yet woefully lacking.

Transforming School Governance

The third major current 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. The 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 education research is to produce reliable evidence about instructional interventions, or “instructional regimes” (Cohen, Raudenbush, & Ball, 2003). Giving educators the flexibility they need to adopt effective practices is an admirable aim, but knowledge of which practices are effective, how to coordinate them schoolwide, and how to conduct required professional development are essential if this kind of reform is to affect teaching and learning in powerful ways.

Conclusion

In sum, policymakers could not directly intervene to improve instruction even if they knew how to 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 science and neuroscience—are not terrifically important and, ultimately, potentially useful for policy and practice. It is to argue, however, that within the domain of applied education research (i.e., 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 article 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.

Questions and Methods

I have argued that instructional improvement is central to the nation’s applied education 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 relied upon to produce good effects over a comparatively broad range of conditions for well-defined target populations at reasonable cost. 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, let us also identify the methodological approaches entailed in this effort. Within this scenario, what is the contribution of mixed methods? Specifically, how might various data collection efforts and analytic strategies combine to accomplish this goal? ¹

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 (Boruch & Mosteller, 2001). 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 the 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. Holland, 1986; Rubin, 1978), 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 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 subpopulation, *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* (Holland, 1986; Rubin, 1974, 1978). This means, in part, that the children assigned to receive a new experimental intervention would have displayed the same average outcome as the control-group children did, had the 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 ensures 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 50% probability of receiving the experimental intervention. Thus a 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. Stated 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.

Although 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, 2001). This means that evaluators of educational innovations have had to rely on nonexperimental methods of attempting to satisfy the assumption of

ignorable treatment assignment. These include, prominently, quasi-experimental designs (Shadish, Cook, & Campbell, 2001) combined with statistical control for potentially confounding variables (c.f. Rosenbaum & 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 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 variables.

In light of such painful experiences, it is not surprising that education evaluators using nonexperimental 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 aiming to increase achievement has been to ensure 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 pretest 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. Stated 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 pretreatment characteristics of children, classrooms, and schools that predispose children to *benefit from the treatment*. This requirement poses a major challenge to valid inference, especially in cases where agents such as administrators, teachers, parents, or even the children themselves select which treatment the children will experience. These 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, pretreatment 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 that something important can be learned about how to improve teaching and learning, when the study does not threaten their basic interests, and when researchers are committed to providing useful information to participating educators and parents.

The recent experience of Robert Slavin and colleagues (Slavin & Borman, 2004) is instructive. Initial attempts to recruit schools to participate in a randomized study of the Success for All program were fruitless. School leaders resisted participating in a study that might assign their school to the control group. Their concern was sensible in light of current pressure on schools to improve under NCLB. So Slavin redesigned the study. In half of the schools, Success for All would be implemented in kindergarten during Year 1 of the study, in kindergarten and first grade during Year 2, and in kindergarten through second grade 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 3 years of the study. Ultimately, all schools would receive the program at all grade levels. But during the first 3 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 ensure 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 ensure the success of this research agenda. I now consider the complementary studies and appropriate research methods needed to ensure success.

Why Experiments Are Not Sufficient

The proposed research is constructed to study these interventions in order to learn what works, so that education 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 dis-

cern the causal effects of such interventions, it may, therefore, appear that a well-planned sequence of randomized experiments would suffice to achieve these aims. Although this simple reasoning has a surface appeal, I reject it. Well-designed randomized experiments are, I believe, necessary but not sufficient for determining what works.

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 *children* 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 the interventions that hold the greatest promise in achieving those outcomes. We need good data to decide which children to target in which settings, because children with particular needs are of the greatest importance in particular instructional situations. For example, third graders who have failed to respond well to good reading instruction likely will 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 for changing the most important outcomes for target children in settings of interest. This process involves a series of well-designed descriptive and correlational studies using a variety of methods, without which the program of randomized experimentation is doomed to fail.

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 settings of classrooms and schools, this might be a good idea. However, in the world as it exists, large-scale intervention studies 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 ensure 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.

Next, I sketch some of the 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. The National Assessment of Educational Programs (NAEP) and the Third International Mathematics and Science Study (TIMSS) 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, policymakers and researchers would not be clear on which outcomes for which children 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 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 themselves constitute an ambitious

program of research that requires a mix of qualitative and quantitative inquiry as we study how students respond to new tasks and how they make meaning of potential survey 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 the 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 to evaluate them with randomized experiments. By consulting expert knowledge, attempting to implement novel programs on small scales, and making preliminary (nonrandomized) assessments, we can determine which interventions to discard or refine before trying large-scale summative tests of effectiveness. 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 is not 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 prerequisite for the construction of a large-scale randomized field trial.

An example of secondary analysis informing intervention comes from TIMSS. Schmidt, McKnight, and Raizen (1997) found that, compared with mathematics

instruction in the United States, 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 United States. 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 turned out that the children in nations that pursued more focused and coherent instruction did substantially better on the mathematics assessments than did U.S. children.

Preliminary research on this subject might also involve secondary analysis of U.S. databases combined with case studies to see whether some U.S. teachers have pursued focused and coherent instruction (as defined by Schmidt et al. using TIMMS) and whether students in those classes fared well. Together, these findings, culled from a variety of descriptive methodologies, both quantitative and qualitative, would 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 U.S. 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 children 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 children lack good teachers or other resources, in which case equalizing resources might be a promising strategy. Alternatively, such children may thrive in

instructional environments that are not effective for other children. In that 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 for answering 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 children 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 U.S. Department of Education, the National Science Foundation, and 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 ensure 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, although necessary, is far from sufficient to achieve the goal of learning about “what works.” Research using a variety of methods is essential and should include

1. Defining the student outcomes that we seek, so that we can change, build, and validate assessments of those outcomes;
2. Supporting 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;

3. Clarifying the subsets of children who are in greatest need of intervention or who are most likely to benefit from new ideas about teaching and learning; and
4. Studying how resource constraints affect the outcomes of interventions, with the aim of ensuring that new approaches are cost effective.

A final goal is to study why an intervention works, why it works for some children and not others, or why it fails. A variety of methodological strategies, including 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” (i.e., settings in which a new intervention is initially found effective) can be compared with 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 Multi-Methods Research

The success of an ambitious program of innovation and experimentation appears to depend on a fairly complex set of interrelated 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 effects 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 education 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 that research managers and funding agencies keep in mind the broad aim—to improve achievement by improving teaching and learning in classrooms—while understanding the complementary efforts required for success.

Finally, the effort thus sketched requires adequate funding. By now, the small fraction of all funding for education that supports education research is well known. It is hard to imagine how an ambitious agenda of randomized trials supported by multidisciplinary and multi-methodological efforts can succeed at present funding levels. The weak yield of applied research has undermined the case for generous funding of education research. On a more hopeful note, successful efforts to produce sound evidence about how to intervene should, in principle, generate wider support for education research, thereby increasing capacity to mount an increasingly ambitious, effective, and diverse research effort.

NOTES

This article was commissioned as background for the forum “Applying Multiple Social Science Research Methods to Educational Problems,” held on December 14, 2004, in Washington, DC. The Center for Education of the National Research Council convened the forum with support from the American Educational Research Association, the American Psychological Association, and the National Science Foundation. Opinions and statements included in the article 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.

¹ Some readers will disagree, of course, about my reading of the research priorities. Yet I 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 dis-

cussion will be empty or—what is worse—misleading, as question-free methodological discussions typically are. 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.

REFERENCES

- Borko, H. (2004). Professional development and teacher learning: Mapping the terrain. *Educational Researcher*, 33(8), 3–15.
- Borman, G. D., Hewes, G. M., Overman, L. T., & Brown, S. (2003). Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research*, 73, 125–230.
- Boruch, R., & Mosteller, F. (Eds.). (2001). *Education, evaluation, and randomized trials*. Washington, DC: Brookings Institution Press.
- Cohen, D. K., Raudenbush, S. W., & Ball, D. L. (2003). Resources, instruction, and research. *Educational Evaluation and Policy Analysis*, 25(2), 1–24.
- Coleman, J. S., Campbell, E., Hobson, C., McPartland, J., Mood, A., Weinfeld, E., et al. (1966). *Equality of educational opportunity*. Washington, DC: U.S. Government Printing Office.
- 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 & F. Mosteller (Eds.), *Education, evaluation, and randomized trials*. Washington, DC: Brookings Institution Press.
- Finn, J. D., & Achilles, C. M. (1990). Answers and 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.
- 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.
- Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396), 945–960.
- Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher*, 33(7), 14–26.

- Krueger, A., & 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.
- Nye, B., Hedges, L. V., & Konstantopoulos, T. (2000). The effects of small classes on achievement: The results of the Tennessee class size experiment. *American Educational Research Journal*, 37, 123–151.
- Rosenbaum, P., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41–55.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66, 688–701.
- Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, 6, 34–58.
- Schmidt, W. H., McKnight, C. C., & Raizen, S. A. (1997). *A splintered vision: An investigation of U.S. science and mathematics education*. Boston: Kluwer Academic.
- Shadish, W. R., Campbell, D. T., & Cook, T. D. (2001). *Experimental and quasi-experimental designs for research*. Boston: Houghton Mifflin.
- Slavin, R., & Borman, G. (2004, July). Preliminary results of the experimental evaluation of Success for All. Paper presented at the national conference Design and Analysis of Group-Randomized Experiments, sponsored by the William T. Grant Foundation and the National Center for Research on Poverty, University of Michigan.

AUTHOR

STEPHEN W. RAUDENBUSH is a Professor in the School of Education and the Survey Research Center, and (by courtesy) the Departments of Statistics and Sociology, at the University of Michigan. He can be contacted at the School of Education, University of Michigan, 610 East University Avenue, Ann Arbor, MI 48109; rauden@umich.edu. His areas of specialization include statistical methods for studying school and classroom effects on student learning; design of surveys and experiments; and applications of hierarchical statistical models in social science research.

Manuscript received January 11, 2005

Revision received April 18, 2005

Accepted April 22, 2005