Sciencephobia

Why education researchers reject randomized experiments

by THOMAS D. COOK

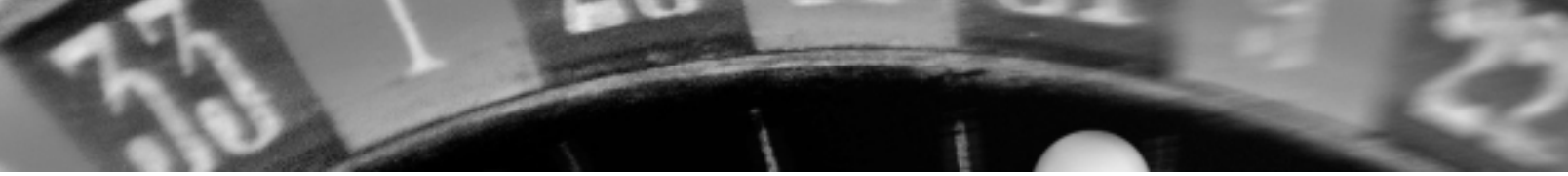
THE AMERICAN EDUCATION SYSTEM, UNIQUELY DECENTRALIZED AMONG INDUSTRIAL nations, has beencontinually roiled by tides of local experimentation, especially during the past 20 years. The spread of whole-school reform models such as Success for All; the imposition of standards and high-stakes tests; the lowering of class sizes and slicing of schools into smaller, independent academies; the explosion of charter schools and push for school vouchers—all these reforms signal a vibrantly democratic school system.

Experimentation, however, means more than simply changing the way we do things. It also means systematically evaluating these alternatives. To scholars, experimentation further suggests: 1) conducting studies in laboratories where external factors can be controlled in order to relate cause more directly to effect; or 2) randomly choosing which schools, classrooms, or students will be exposed to a reform and which will be exposed to the alternative with which the reform is to be compared. When well executed, random assignment serves to rule out the possibility that any post-reform differences observed between the treatment and control groups are actually due to pre-existing differences between the two groups rather than to the effects of the reform. The superiority of random assignment for drawing conclusions about cause and effect in nonlaboratory settings is routinely recognized in both the philosophy of science literature and in methods texts in health, public health, agriculture, statistics, microeconomics, psychology, and those parts of political science and sociology that deal with improving the assessment of public opinion.

Since most education research must take place in actual school settings, random assignment would seem to be a highly appropriate research tool. However, though the American education system prizes experimentation in the sense of trying many new things, it does not promote experimentation in the sense of using random assignment to assess how effective these new things are. One review showed that not even 1 percent of dissertations in education or of the studies archived in ERIC Abstracts involved randomized experiments. A casual review of back issues of the premier journals in the field, such as the *American Educational Research Journal* or *Educational Evaluation and Policy Analysis*, tells a similar story. Responding to my query, a nationally recognized colleague who designs and evaluates curricula replied that in her area randomized experiments are extremely rare, adding, "You can't get districts to randomize or partially adopt after a short pilot phase because all parents would be outraged."

Very few of the major reform proposals currently on the national agenda have been subjected to experimental scrutiny. I know of no randomized evaluations of standards setting. The "effective schools" literature includes no experiments in which the supposedly effective school practices were randomly used in

PHOTOGRAPH BY IMAGE 100



some schools and withheld from others. Recent studies of whole-school-reform programs and school management have included only two randomized experiments, both on James Comer's School Development Program, which means that the effects of Catholic schools, Henry Levin's Accelerated Schools program, or Total Quality Management have never been investigated using experimental techniques. School vouchers are a partial exception to the rule; attempts have been made to evaluate one publicly funded and three privately funded programs using randomized experiments. Charter schools, however, have yet to be subjected to this method. On smaller class sizes, I know of six experiments, the most recent and best known being the Tennessee class-size study. On smaller schools I know of only one randomized experiment, currently under way. In fact, most of what we know about education reforms currently depends on research methods that fall short of the technical standard used in other fields.

Equally striking is that, of the few randomized experiments cited above, nearly all were conducted by scholars whose training is outside the field of education. Educators Jeremy Finn and Charles Achilles began the best-known class-size experiment, but statisticians Frederick Mosteller, Richard Light, and Jason Sachs popularized the study, and economist Alan Krueger has conducted an important secondary analysis. Political scientist John Witte conducted the Milwaukee voucher study, while political scientists Jay Greene and his colleagues and economist Cecelia Rouse reanalyzed the data. Sociologists and psychologists conducted the Comer studies. Economists James Kemple and JoAnn Leah Rock are running the ongoing experiment on academies within high schools. Political scientist William Howell and his colleagues did the work on school-choice programs in Washington, D.C.; New York City; and Dayton, Ohio. Scholars with appointments in schools of education, where we might expect the strongest evaluations of school reform to be performed, evidence a 20-year near-drought when it comes to randomized experiments.

Such distaste for experiments contrasts sharply with the practices of scholars who do school-based empirical work but don't operate out of a school of education. Foremost among these are scholars who research ways to improve the mental health of students or to prevent violence or the use of tobacco, drugs, and alcohol. These researchers usually have disciplinary backgrounds in psychology or public health, and they routinely assign schools or classrooms to treatments randomly. Randomized experiments are commonplace in some areas of contemporary research on primary and secondary schools. They're just not being done by researchers who were trained in education schools.

Where the Research Dollars Flow

Of 84 program evaluations and studies planned by the Department of Education for fiscal year 2000, just one involved a randomized field trial.

Purpose of the study	Number
Randomized field trial	**1**
Survey of need	51
Program implementation/ monitoring	49
Non-randomized impact evaluation	15
Total	116*

**Studies could have more than one primary purpose.*

SOURCE: Robert Boruch, Dorothy de Moya, and Brooke Synder, in Robert Boruch and Frederick Mosteller, eds., *Evidence Matters* (Brookings, 2001).

Dealing with Complexity

In schools of education, the intellectual culture of evaluation actively rejects random assignment in favor of alternatives that the larger research community has judged to be technically inferior. Education researchers believe in a set of mutually reinforcing ideas that provides what for them is an overwhelming rationale for rejecting experiments on any number of philosophical, practical, or ethical grounds. Any Ph.D. from a school of education who was exposed to the relevant literature on evaluation methods has encountered arguments against experiments that appeared cogent and comprehensive. For older education researchers, all calls to conduct formal experiments probably have a "déja vu" quality, reminding them of a battle they thought they had won long ago—the battle against a "positivist" view of science that privileges the randomized experiment and its related research and development model whose origins lie in agriculture, health, public health, marketing, or even studies of the military. Education researchers consider this model irrelevant to the special organizational complexity of schools. They prefer an R&D model based on various forms of management consulting.

In management consulting, the crucial assumptions are that 1) each organization possesses a unique culture and set of goals; therefore, the same intervention is likely to elicit different results depending on a school's history, organization, personnel, and politics; and 2) suggestions for change should creatively blend knowledge from many different sources—from general organizational theories, from deep insight into the district or schools under study, and from "craft" knowledge of what is likely to improve schools or districts with particular characteristics. Scientific knowledge about effectiveness is not particularly prized in the management-consulting model, especially if it is developed in settings different from those where the knowledge is to be applied.

As a central tool of science, random assignment is seen as the core of an inappropriate worldview that obscures each school's uniqueness, that oversimplifies the complicated nature of cause and effect in a school setting, and that is naive about the ways in which social science is used in policy debates. Most education evaluators see themselves as the vanguard of a post-positivist, democratic, and craft-based model of knowledge growth that is superior to the elitist scientific model that, they believe, has failed to create useful and valid knowledge about improving schools. Of the reasons critics articulate for rejecting random

assignment as an evaluation tool, some are not very credible, but others are and should inform the design of future studies that use random assignment. Let's deal with some of the major objections in turn.

The world is ordered more complexly than a causal connection from A to B can possibly capture. For any given outcome, randomized experiments test the influence of only a few potential causes, often only one. At their most elegant, they can responsibly test only a modest number of interactions between different treatments or between any one treatment and individual differences at the school, classroom, or individual level. Thus, randomized experiments are best when the question of causation is simple and sharply focused.

Lee Cronbach, perhaps the most distinguished theorist of educational evaluation today, argues that in the real world of education too many factors influence the system to isolate the one or two that were the primary causes of an observed change. He cannot imagine an education reform that fully explains an outcome; at most there will be just one cause of any change in this outcome. Nor can he imagine an intervention so general in its effects that the size of a cause-effect relationship remains constant across different populations of students and teachers, across different kinds of schools, across the entire range of relevant outcomes, and across all time periods. Experiments cannot faithfully represent a real world characterized by multivariate, nonlinear (and often reciprocal) causal relationships Moreover, few education researchers have much difficulty detailing contingencies likely to limit the effectiveness of a proposed reform that were never part of a study's design.

There is substance to the notion that randomized experiments speak to a simple, and possibly oversimplified, theory of causation. However, many education researchers speak and write as though they accept certain contingency-free causal connections—for example, that small schools are better than large ones; that time on task raises achievement; that summer school raises test scores; that school desegregation hardly affects achievement; and that assigning and grading homework improves achievement. They also seem to be willing to accept some propositions with highly circumscribed causal contingency—for instance, that reducing class size increases achievement (provided that it is a "sizable" change and that the reduction is to fewer than 20 students per class); that Catholic schools are superior to public ones in the inner-city but not in suburban settings. Commitment to a full explanatory theory of causation has not precluded some education researchers from acting as if very specific interventions have direct and immediate effects.

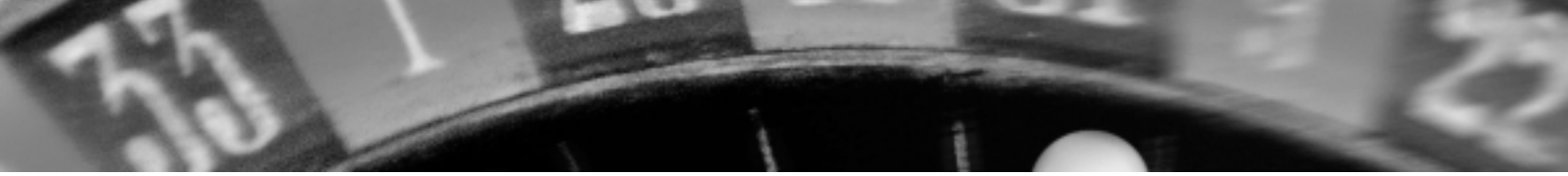
Quantitative research has been tried and has failed. Education researchers were at the forefront of the flurry of social experimentation that took place at the end of the 1960s and through the 1970s. Quantitative studies of Head Start, Project Follow Through, and Title I concluded that, for all three programs, there were no replicable effects of any magnitude that persisted over time. Such results provoked hot disputes over the methods used, and many educational evaluators concluded that quantitative evaluation of all kinds had failed. Some evaluators turned to other methods of educational evaluation. Others turned to the study of school management and program implementation in the belief that poor management and incomplete implementation explained the disappointing results. In any event, dissatisfaction with quantitative evaluation methods grew.

However, none of the most heavily criticized quantitative studies involved random assignment. I know of only three randomized experiments on education reform available at the time. One was of the second year of "Sesame Street," where cable capacity was randomly assigned to homes in order to promote differences in children's opportunity to view the show. The second experiment was the widely known Perry Preschool Project in Ypsilanti, Michigan. The third involved only 12 youngsters who were randomly assigned to a desegregated school. Only the desegregation study involved primary or secondary schools. Thus it was not accurate to claim in the 1970s that randomized experiments had been tried and had failed. Only nonexperimental quantitative studies had been done, and few of these would pass muster today as even high-quality *quasi*-experiments.

Random assignment is not politically, administratively, or ethically feasible in education. The small number of randomized experiments in education may reflect not researchers' distaste for them but a simple calculation of how difficult they are to mount in the complex organizational context of schools. School district officials do not like the focused inequities in school structures or resources that random assignment usually generates, fearing backlash from parents and school staff. They prefer it when individual schools can choose which reforms they will implement or when changes are made on a district-wide basis. Some school staff members also have administrative concerns about disrupting routines and ethical concerns about withholding potentially helpful treatments from students and teachers in need.

Of the few randomized experiments in education, nearly all were conducted by scholars whose training is outside the field of education.

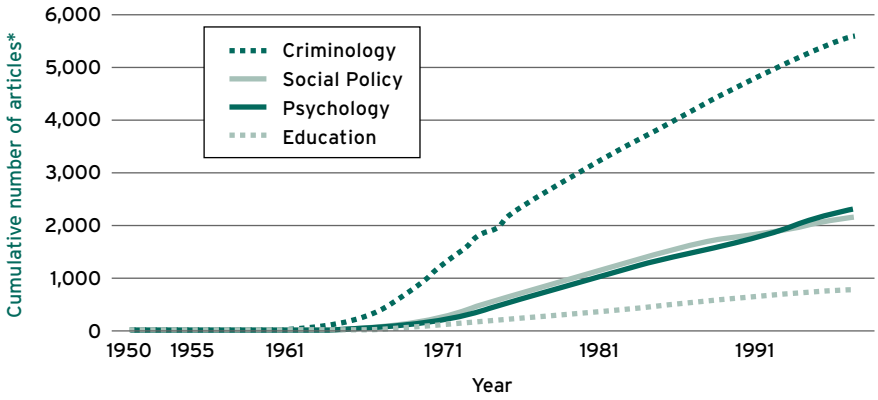
Surely it is not easy to implement randomized experiments of school reform. In many of the recent experiments, schools have dropped out of the experiment in different proportions, often because a new principal wanted to change what his predecessor had recently done, including eliminating the reform under study.



Then there are the cases of possible treatment crossover, as happened in one of my own studies in Prince George's County, Maryland. One principal in an experimental school was married to someone teaching in a control school, and they discussed their professional life at home; one control principal really liked the reform under study and tried to bring parts of it to his school; and the daughter of one of the program's senior officials taught in a control school. In a similar vein, the Tennessee class-size experiment compared classrooms within the same schools. What did Tennessee teachers in the larger classes make of the situation whereby some colleagues in the same school taught smaller classes at the same grade level? Were they dispirited enough to work less? To avoid such possibilities, most public-health evaluations of prevention programs (such as those aimed at reducing drug use) use comparisons between schools instead of between classrooms within the same school. All randomized experiments in education have to struggle with issues like these.

Education Lags Behind (Figure 1)

While the total number of articles about randomized field trials in other areas of social-science research has steadily grown, the number in education research has trailed behind.



* Articles about definite and possible randomized field trials.

SOURCE: Robert Boruch, Dorothy de Moya, and Brooke Snyder, 2001

What does it take to mount randomized experiments? Political will plays an important role. In the health sciences, random assignment is common because it is institutionally supported by funding agencies and publishing outlets and is culturally supported through graduate training programs and the broadly accepted practice of clinical trials. Public-health researchers have learned to place a high priority on clear causal inferences, a priority reinforced by their funders (mostly the National Institutes of Health, the Centers for Disease Control, and the Robert Wood Johnson Foundation). The health-related studies conducted in schools tap into this institutional and cultural structure. Similar forces operate with the rapidly growing number of studies of preschool education that use random assignment. Most are the product of congressional requirements to assign at random; the high political and scholarly visibility of the Perry Preschool and Abecedarian projects that used random assignment; and the involvement of researchers trained in psychology and microeconomics, fields where random assignment is valued.

Contrast this with educational evaluation. Reports from the Department of Education's Office of Educational Research and Improvement (OERI) are supposed to detail what is known to work. However, neither the work of educational historian Maris Vinovskis nor my own reading of OERI reports suggests that any privilege is being accorded to random assignment. At a recent foundation meeting on teaching and learning, a representative of nine regional governors discussed the lists of best practices that are being widely disseminated. He did not care, and he believed that the governors do not care, about the technical quality of the designs generating these lists; the major concern is that educators can deliver a *consensus* on each practice. When asked how many of these best practices depended on randomized experiments, he guessed it would be close to zero. Several nationally known education researchers were present. They too replied that random assignments probably played no role in generating these best-practice lists. No one seemed to feel any distress at this.

Random assignment is premature because it assumes conditions that do not yet pertain in education. As the research emphasis shifted in the 1970s to understanding schools as complex social organizations with severe organizational problems, randomized experiments must have seemed premature. A more pressing need was to understand management and implementation, and to this end, more and more political scientists and sociologists of organizations were recruited into schools of education. They brought with them their own strongly held preference for qualitative methods and their memories of the wars between quantitative and qualitative methods in their own disciplines.

However, school research need not be predicated only on the idea of schools as complex organizations. Schools were once conceptualized as the physical structure containing many self-contained classrooms in which teachers tried to deliver effective curricula using instructional practices that demonstrably enhance students' academic performance. This approach privileged curriculum design and instructional practice over the schoolwide factors that have come to dominate understandings of schools as complex organizations—factors like strong leadership, clear and supportive links to the world outside of school, a building-wide community focused on learning, and the pursuit of multiple forms of professional development.

Many important consequences have flowed from the intel-

lectual shift in how schools are conceptualized. One is the lesser profile accorded to curriculum and instructional practice and to what happens once the teacher closes the classroom door; another is the view that random assignment is premature, given its dependence on expert school management and high-quality program implementation; and another is the view that quantitative techniques have only marginal usefulness for understanding schools, since a school's governance, culture, and management are best understood through intensive case studies.

However, the aim of experiments is not to explain all sources of variation; it is to probe whether the school reform idea makes a difference at the margin, despite whatever variation exists among schools, teachers, students, or other factors. It is not an argument against random assignment to claim that some schools are chaotic, that implementation of a reform is usually highly variable, and that treatments are not completely faithful to their underlying theories. Random assignment does not need to be postponed while we learn more about school management and implementation.

Nonetheless, the more we know about these matters, the better we can randomize and the more management and implementation issues can be worthy objects of study *within experiments*. Advocates of random assignment will not be credible in educational circles if they assume that reforms will be implemented uniformly. Experimenters need to be forthright that school-level variation in implementation quality will often be very large. It is not altogether clear that schools are more complex than other settings where experiments are routinely done—say, hospitals—but most school researchers seem to believe this, and it seems a reasonable working assumption.

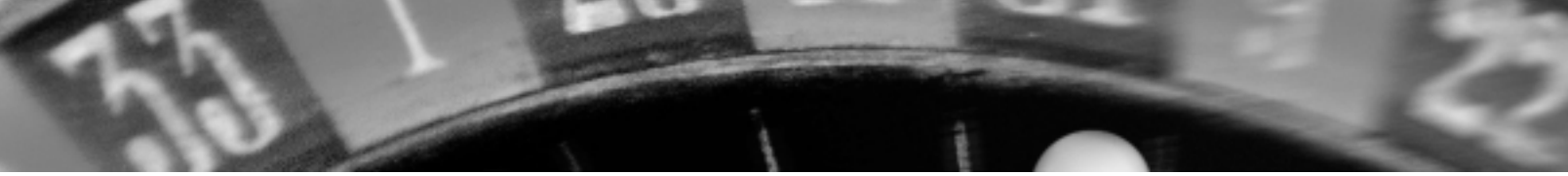
Thirty years after vouchers were proposed, we still have no clear answer about them. Thirty years after James Comer began his work that has resulted in the School Development Program, and again we have no clear answer. Almost 20 years after Henry Levin began Accelerated Schools; here too we have no answer. While premature experimentation is indeed a danger, these time lines are inexcusable. The federal Obey-Porter educational legislation cites Comer's program as a proven program worth replicating elsewhere and provides funds for this. But when the legislation passed, the only available evidence about the program consisted of testimony; a dozen or so empirical studies by the program's own staff that used primitive quasi-experimental designs; and the most-cited single study confounded the court-ordered introduction of the program with a simultaneously ordered reduction in class sizes of 40 percent. To be restricted to such evidence when making a decision about federal funding verges on the irresponsible.

Unlike medicine or public health, education has no tradition of multisite experiments with national reach. Single experiments of unclear reach, done only in Milwaukee, Washington, Chicago, and Tennessee, are what we typically find. Moreover, some kinds of school reform have no fixed protocol, and it is possible to imagine implementing vouchers, charter schools, or programs like Comer's or Total Quality Management schools in many different ways. Indeed, the Comer programs in Prince George's County, Chicago, and Detroit are different from one another in many major specifics. The nonstandardization of many treatments requires even larger samples than those typically used in medicine and public health. Getting cooperation from so many schools is not easy, given the history of local control in education and the absence of a tradition of random assignment. Still, larger individual experiments can be conducted than are being done today.

Random assignment is not needed because there are other less irritating methods for generating knowledge about cause and effect. Most researchers who evaluate education reforms believe there are superior alternatives to the randomized experiment. These methods are superior, they believe, because they are more acceptable to school personnel, because the knowledge they generate reduces enough uncertainty about causation to be useful, because the knowledge is relevant to a broader array of important issues than merely identifying a causal connection, and because schools are especially likely to use the results for self-improvement. No single alternative is universally recommended, and here I'll discuss only two: intensive qualitative case studies and quasi-experiments.

- *Intensive case studies.* Cronbach asserted that the appropriate methods for educational evaluation are those of the historian, journalist, and ethnographer, not the scientist. Most educational evaluators now seem to prefer case-study methods for learning about reforms. They believe that these methods are superior because schools are less squeamish about allowing ethnographers through the door than experimentalists. They also believe that qualitative studies are more flexible. They provide simultaneous feedback on the many different kinds of issues worth raising about a reform—issues about the quality of implementation, the meaning various actors ascribe to the reform, the primary and secondary effects of the reform, its unanticipated side effects, and how different subgroups of teachers and students are affected. Entailed here is a flexibility of purposes that the randomized experiment cannot match, given its limited central purpose of facilitating clear causal inference.

It is not altogether clear that schools are more complex than other settings where experiments are routinely done—say, hospitals.



A further benefit relates to schools actually using the results. Ethnography *requires* attention to the unfolding of results at different stages in a program's implementation, thus generating details that can be fed back to school personnel and that also help explain *why* a program is effective. A crucial assumption is that school staff are especially likely to use a study's results because they have a better ongoing relationship with qualitative researchers than they would have with quantitative ones. Of course, the use in question is highly local, often specific to a single school, while the usual aspiration for experiments is to guide policy changes that will affect large numbers of districts and schools.

The downside of case studies is the question of whether this process reduces enough uncertainty about causation to be useful. With qualitative methods it is difficult to know just how the group under study would have changed had the reform not been in place. The rationale for preferring an experiment over an intensive case study has to be the value of a clear causal inference, of not being wrong with the claim that a reform is effective or not. Of course, one can have one's cake and eat it too, for there are no compelling reasons why case study methods cannot be used within an experiment to extend its reach. While black-box experiments that generate no knowledge of process may be common, they are not particularly desirable. Nor are they the only kinds of experiments possible.

- *Quasi-experiments.* Quasi-experiments are like randomized experiments in purpose and in most of their structural details. The defining difference is the absence of random assignment and hence of a demonstrably valid causal counterfactual. The essence of quasi-experimentation is the search, more through design than statistical adjustment, to create the best possible approximation of this missing counterfactual. However, quasi-experiments are second best to randomized experiments in the clarity of causal conclusions. In some quarters, quasi-experiment has come to connote any study that is not an experiment or any study that includes some type of nonequivalent control group or pretreatment observation. Indeed, many of the studies calling themselves quasi-experiments in educational evaluation are of types that theorists of quasi-experimentation reject as usually inadequate. To judge by the quality of the educational evaluation work I know best—on school desegregation, Comer's School Development Program, and bilingual education—the average quasi-experiment in these fields inspires little confidence in its conclusions about effectiveness. Recent advances in the design and analysis of quasi-experiments are not getting into research evaluating education.

The average quasi-experiment in education inspires little confidence in its conclusions about effectiveness.

Moving Forward

It will be difficult to persuade the current community of educational evaluators to begin doing randomized experiments solely by informing them of the advantages of this technique, by providing them with lists of successfully completed experiments, by telling them about new methods for implementing randomization, by exposing them to critiques of the alternative methods they prefer, and by having prestigious persons and institutions outside of education recommend that experiments be done. The research community concerned with evaluating education reforms is a community in which all parties share at least some of the beliefs outlined above. They are convinced that anyone pursuing a scientific model of knowledge growth is an out-of-date positivist seeking to resuscitate debates that are rightly dead.

Some rapprochement might be possible. At a minimum, it would require advocates of experimentation to be explicit about the real limits of their preferred technique, to engage their critics in open dialogue about the critics' objections to randomization, and to assert that experiments will be improved by paying greater attention to program theory, implementation specifics, quantitative *and* qualitative data collection, causal contingency, and the management needs of school personnel as well as of central decisionmakers.

Though it is desirable to enlist the current community of educational evaluation specialists in supporting randomized experiments, it is not necessary to do so. They are not part of the tiny flurry of controlled experiments now occurring in schools. Moreover, in several substantive areas Congress has shown its willingness to mandate carrying out controlled studies, especially in early-childhood education and job training. Therefore, end runs around the education research community are conceivable. This suggests that future experiments could be carried out by contract research firms, by university faculty members with a policy science background, or by education faculty who are now lying fallow. It would be a shame if this occurred and restricted our access to those researchers who know best about micro-level school processes, about school management, about how school reforms are actually implemented, and about how school, state, and federal officials tend to use education research. It would be counterproductive for outsiders to school-reform research to learn anew the craft knowledge insiders already enjoy. Such knowledge genuinely complements controlled experiments.

–Thomas D. Cook is a professor of sociology, psychology, education, and social policy at Northwestern University. This article is adapted from a chapter that will appear in Evidence Matters *(Brookings, forthcoming). To view his essay in its entirety, log on to www.educationnext.org.*