Can the Federal Government Improve Education Research?

BRIAN JACOB and JENS LUDWIG

Recent dissatisfaction with public education in the United States has been matched by dismay with the current state of education research. A common complaint is that education research is good at description and hypothesis generation but not at answering causal questions about the effects of education policies on student outcomes. In this vein, many policymakers have expressed frustration that, as Ellen Condliffe Lagemann has noted, “education research has not yielded dramatic improvements in practice of the kind one can point to in medicine.” Such dissatisfaction has contributed to a number of recent federal policy changes intended to improve the quality of research in education, including the creation of a new Institute of Education Sciences (IES) to support increased experimentation within education and an emphasis on the use of teaching methods supported by “scientifically-based research” in the 2001 No Child Left Behind Act (NCLB).

In this paper we consider the possible effects of these recent changes on the state of education research. We focus on what might be termed program or policy evaluation—research that aims to support causal inferences about the efficacy of specific educational programs or policies. Examples include studies that examine whether smaller class size improves student achievement, whether a particular reading curriculum leads to increased reading comprehension, and whether “pull-out” programs are more effective than “push-in” programs for students with learning disabilities. It is important to note that a great deal of research in education does not aim to answer these types of questions but rather

Thanks to Jennifer Brya for excellent research assistance and to Robert Boruch, Phoebe Cottingham, Steven Glazerman, Robin Jacob, Diane Ravitch, Cecilia Rouse, Maris Vinovskis, and various Brookings conference participants for helpful comments.
addresses the processes underlying teaching and learning or human development more generally. Although obviously this “basic” research can provide insights that directly benefit the development of more effective curriculum and pedagogy, our interest here is on the type of evaluation research designed to inform larger policy decisions relevant to education reform (though for convenience we sometimes refer generically to “education research”).

Understanding how recent policy changes may affect the state of education policy research and, ultimately, education practice requires some consideration of why the quality of past research has, on average, been so poor. Over the past several years, growing attention has been devoted to the question of why randomized field trials, considered the gold standard for causal inference in medical research and most of the social sciences, are not used more often in education. Candidate explanations include the practical and political difficulties associated with launching experiments in education and problems confronting the federal agencies that support education research. Left unanswered are several additional questions that are relevant in considering the effects of recent policy changes on education research. Why is the average quality of nonexperimental evaluation research in education so low? Why is much of the education evaluation work that is supported by private foundations and other nongovernmental sources of such low quality? More generally, why has “good” evidence from randomized experiments (or credible natural or quasi experiments) not driven out “bad” evidence over time?

In what follows, we suggest that the supply of high-quality education policy research is limited because the demand for it is limited. The supply of high-quality research is likely to be greatest in markets in which there is significant demand for the end product itself and good research can be distinguished from bad. Neither of these conditions currently seems to hold with much force in education because of “market failure” in the markets for both educational outcomes and education research.

In the past, there has been little demand for educational outcomes—specifically, academic achievement—among parents, educators, and policymakers. Possible explanations for this state of affairs include the difficulty parents have in determining the value added of their child’s school, parental preferences for nonacademic as well as academic outcomes from schools, and a traditional lack of accountability within the public schools, which reduces the incentives facing educators to demand high achievement from their students. Insofar as high-quality education research is useful primarily as a means to improve student achievement most efficiently, low demand for achievement will necessarily translate into low demand for high-quality research.
Moreover, information problems in the market for education research have led to the undersupply of high-quality research. Typically, the costs of deriving evidence from randomized field trials or well-crafted quasi-experimental studies exceed those associated with other, nonexperimental, research designs. Yet research consumers—educators, policymakers, and the public—have limited information about the relative benefits of these different approaches. Even research producers face information problems about product quality. Because the additional costs of randomized trials or high-quality quasi experiments are usually more obvious than the additional benefits, high-cost research designs are used relatively rarely by education policy analysts.

In light of our diagnosis, how likely are recent changes in federal policy to improve the average quality of program evaluation research in education? The most visible change involves the reorganization of the Department of Education, including the creation of the IES. One important part of the institute’s mission is to increase the supply of high-quality evidence in education, particularly evidence derived from randomized experiments. This is a positive step with encouraging results to date, although the ultimate impact on education research may be limited for a number of reasons. Historically, the federal government’s funding support for education research has been quite modest. Moreover, the language in the act that creates the IES is sufficiently ambiguous to allow its future administrators to scale back their commitment to randomized evaluations should they so choose. The commitment of IES administrators to experimentation could be affected in the future by the same bureaucratic and political constraints that have hampered previous attempts to improve the quality of education research supplied by the Department of Education. Moreover, efforts to change the way the Department of Education funds research will have little effect on education research funded by other federal agencies such as the National Science Foundation.

Perhaps more important in the long run are recent efforts to implicitly or explicitly stimulate the demand for research and to fix the information problems that plague the market for education research. For example, the growing tide of state accountability reforms, together with the recent federal NCLB Act, may increase the demand for high-quality research by local education practitioners and policymakers. To the extent to which this occurs, guidance about what the best evidence currently tells us may come from the IES’s new What Works Clearinghouse, operated by the American Institutes for Research and the Campbell Collaboration, whose goal is to “provide educators, policymakers, researchers, and the public with a central and trusted source of scientific evidence of what works in education.”
An alternative approach is suggested by the inclusion of requirements in recent federal legislation that federal funding to schools be spent on methods supported by scientifically based research. This approach essentially requires local educators and policymakers to pay attention to research rather than hope that accountability leads them to realize the benefits of research for practice. The information problem in the research market is solved by the establishment of a minimum quality standard rather than by hoping that research consumers are inclined and able to appropriately weigh the best available evidence (or the best available research synthesis) on a given topic. In principle, this approach could be employed by state as well as federal agencies, which account for a much larger share of total elementary and school spending in the United States and could be expanded to other areas of instruction beyond reading.

These recent demand-side initiatives are particularly promising because they circumvent the budgetary, bureaucratic, and political factors that have limited the impact of previous efforts to increase the supply of high-quality federally funded education research. Previous supply-oriented efforts have also been limited by the difficulty that education evaluation consumers and even producers have in determining the quality of the research that is produced. This same information problem could also limit efforts to stimulate the demand for high-quality education evaluation research.

In our view, the prospects that recent federal policy changes will improve the quality of education evaluation work could be strengthened by improving our understanding of the benefits and costs of different research approaches. The efficient allocation of education research funds by the IES and other federal agencies will require some understanding of whether the additional costs of mounting a randomized experiment to answer a given policy question are justified by the benefits. The relative benefits of randomized experiments will depend in part on the ability of different nonexperimental methods to reproduce experimental findings in a given application. Which nonexperimental research designs are best able to reproduce experimental inferences, and under what conditions? The IES and other agencies should support studies that seek to “evaluate the evaluations” and answer such questions, which in the long run may be at least as important as federal funding for the randomized experiments themselves. Information on the performance of different research designs in different applications is also important for demand-side strategies that seek to steer consumers to the best available research (as does the What Works Clearinghouse) or establish a minimum quality standard for the evaluation work that supports interventions used by federal grantees.
The State of Program Evaluation Research in Education

The goal of program evaluation research in general is to determine the causal effect of a particular program or policy (that is, a treatment). Social scientists define the causal effect of a policy as the difference between the outcomes students experience when the policy is administered and those the students would have experienced had they been exposed to a different policy. Because one can never directly observe a “counterfactual” outcome—if a given child has gone through third grade in a classroom with fifteen other students, we cannot observe what that child’s academic outcomes would have been had he or she gone through third grade in a classroom with twice as many students—it is difficult to construct a valid comparison group.

If, for example, the achievement of students in small classes is compared with that of students in large classes, one might find a positive correlation between small classes and high achievement—students in small classes generally do better than their peers in large classes. However, does this correlation represent a causal effect? There may be other “omitted variables” that are related to both the student’s placement in a small class and his or her outcomes. Suppose that small classes are typically found in more affluent school districts. In this case, the failure to adequately measure the socioeconomic background of students will cause unmeasured aspects of family background to be confounded with the effects of smaller classes. The central challenge in program evaluation research is to distinguish between correlation and causation.

Randomized experiments solve the problem of deriving a valid counterfactual by randomly assigning a population of students (or teachers, or schools) to either a “treatment group,” which is exposed to the policy innovation of interest, or a “control group,” which is exposed to some alternative policy or program. Because of random assignment, the treatment and control groups would, on average, have had identical outcomes had both groups been exposed to the same policy. Therefore, any difference in average outcomes can confidently be attributed to the causal effects of the difference in policies that are applied.

Yet randomized experiments are still relatively rare in education research. Robert Boruch, Dorothy de Moya, and Brooke Snyder, reporting on a search of articles in the American Educational Research Journal from the journal’s founding in 1964 through 1998, find low levels and little increase in the publication of research using randomized field trials. Moreover, as Robert Slavin notes, many of those experiments that are conducted in education provide
“suggestions about how to think about daily teaching problems, not guides to the larger questions educators and policymakers must answer.”9 Much more common in education are qualitative studies and nonexperimental quantitative studies. Qualitative research is often favored in education because of its ability to pay close attention to subtle differences between students, families, teachers, classrooms, and schools that are difficult to capture using larger-scale survey questionnaires. And, indeed, as a descriptive tool that helps to identify new testable hypotheses, qualitative observation is often quite powerful. By trading breadth for depth, however, qualitative research often has limited power to derive valid counterfactual outcomes for the subjects being studied. Moreover, qualitative research is often unable to determine whether a particular relationship, phenomenon, or effect is generalizable to a wider population.

For example, countless newspaper and magazine articles have been devoted to the concern that the gap in the relative academic achievement of whites and African Americans owes to some degree to a withholding of effort among black students, who are reportedly concerned that exertion of effort will be derided by their peers as “acting white.” The evidence for this view comes in large part from a qualitative study of black students attending a disproportionately black urban school as well as from media interviews with minority students who report such anxieties.10 For the “acting white” concern to explain the difference in achievement between blacks and whites, however, either the prevalence or the consequences of pressure not to achieve must be greater for blacks than for whites—an issue that obviously cannot be addressed by studying the experiences of black students alone. More systematic empirical analysis of representative samples of black and white students yields little evidence either that blacks exert less effort in school than whites who come from similar family backgrounds or that the social penalty paid by high achievers is different for blacks and whites.11

Quantitative studies based on large, representative populations provide a stronger basis for identifying general relationships. However, in education such studies often use weak nonexperimental research designs that have difficulty distinguishing between correlation and causation. For example, since 1966, when the Coleman Report was released, countless studies in education have used nonexperimental data to assess the relation between class size and student outcomes. These studies generally regress measures of class size (or student-teacher ratio) against student outcomes, controlling for other observable student, family, and school characteristics in an attempt to compare “otherwise comparable” students in small and large classes. The findings from this literature are quite mixed.12 However, this strategy naturally raises a
question: Why do observationally equivalent students wind up in classes of different sizes or in schools with different student-teacher ratios? It is likely that the unobserved processes or student characteristics that lead apparently similar children to have different educational experiences may themselves influence educational outcomes.

Although in theory they provide the strongest basis for causal inference, randomized field trials, like any research design, have limitations. Unlike laboratory experiments, real-world experiments are difficult to implement with complete fidelity to the idealized experimental design. Consider, for example, the well-known Tennessee STAR (Student/Teacher Achievement Ratio) class-size experiments, which randomly assigned classrooms within schools to a “small” or “large” (or “large with aide”) category. Evaluation of STAR has been complicated by the fact that some parents demanded that their children be transferred from large to small classrooms—that is, there was crossover between the control and treatment groups. Moreover, some of the students who initially enrolled in STAR did not participate in all of the follow-up survey and testing sessions. Selective sample attrition can bias the results of even randomized experiments, as might occur when parents transfer their children to private schools rather than leave them in large public school classrooms. Reanalysis of the STAR data that attempts to adjust for these problems still suggests that small classes, on average, raise student achievement by roughly one-fifth of a standard deviation.13

Social experiments also typically provide limited power for identifying the mechanisms underlying a particular treatment effect. The Tennessee STAR experiments, for example, offer limited insight regarding the factors generating improved student achievement—whether it is increased time spent on task, changes in pedagogy, or the closer teacher-student relationships that occur in small classes. Understanding of these mechanisms may help improve the effectiveness or efficiency of the treatment being evaluated or aid in the design of alternative treatments that achieve the same ends at a lower cost.14

More generally, certain types of programs and policies are more suitable than others for evaluation by randomized trials. Because most randomized field trials are limited in scope, for example, they sometimes cannot provide estimates of the effects of policies that are implemented on a larger scale (what economists refer to as general equilibrium effects). Tennessee STAR did not require participating schools to hire many more teachers. However, when the state of California launched a large-scale effort to reduce class size across the board in the early grades, the resulting increase in teacher vacancies may have contributed to a decline in average teacher quality, particularly in schools.
serving disproportionately minority student bodies. As another example, existing randomized field trials on school vouchers are not informative about the effects of private school competition on public school performance because the number of school vouchers issued through these experiments is small, and so the competitive threat posed to the local public school system is also modest.

The possibility of heterogeneity in treatment effects also complicates efforts to generalize from randomized field trials to large-scale policy effects. For example, the schools that volunteered to participate in STAR might in principle have been those that believed themselves best able to implement class-size reduction. In this case, the effects of class-size reduction in the STAR schools could overstate to some degree the benefits from reducing class size across all public schools within the state.

Another relevant consideration is the unit of analysis for the educational intervention of interest. The STAR experiment involved random assignment of students and teachers to different classroom-level treatment conditions. It is hard to imagine using this type of randomized experimental design to understand, for example, the effects of mayoral control over a city’s public school system, because it would require randomly assigning entire cities to treatment or control groups. For all of these reasons, nonexperimental analysis will inevitably be required to answer some questions that do not readily lend themselves to randomized experimentation.

Perhaps most promising for causal inference when randomized field trials are not possible is the set of nonexperimental research designs often referred to as natural or quasi experiments. Quasi-experimental designs are distinguished from standard regression approaches in that they clearly identify the source of variation in treatment used in the analysis and generally rely on aspects of the intervention assignment process that are, if not quite random, then arguably unrelated to other unobserved student, parent, or school characteristics that directly influence the outcome of interest.

Some natural experiments are generated when states, municipalities, or school systems implement sharp changes in practices or policies that apply to some students but not to others. For example, Chicago adopted a comprehensive accountability reform in 1996 that caused a pronounced divergence in test score trends compared with other large midwestern cities over the same period of time. This type of “difference in difference” research design will yield valid estimates for the causal effect of the education program or policy of interest only in cases in which the jurisdictions that are being compared would have had similar trends in test scores in the absence of the “treatment” (in this case, accountability reform). In the case of accountability changes in
Chicago, that the two sets of cities had similar test score trends even before Chicago’s reform was implemented provides some evidence in support of this identifying assumption.\textsuperscript{18} 

In other cases, natural experiments are generated by cutoffs in school assignment policies. For example, estimating the causal effect on students of summer school and grade retention has long been complicated by the fact that students subject to these treatments are more educationally at risk than their counterparts. In earlier work, Brian Jacob and Lars Lefgren overcome this problem by noting that in Chicago, summer school assignments and grade retention decisions were made on the basis of students’ scores on a standardized achievement test.\textsuperscript{19} Students scoring just above a given threshold are likely to be quite similar in almost all respects to those who score just below the cutoff, except that the former will not be subject to summer school and grade retention. Comparing the outcomes of students on either side of the cutoff suggests that both summer school and grade retention produce at least short-term gains in student test scores. In principle, this type of “regression discontinuity”\textsuperscript{20} could be widely used in education research, given that school systems often use hard-and-fast cutoff points for a number of educational decisions.

**The Demand for Education Evaluation Research**

Previous discussions about the quality of education research have focused mainly on supply-side issues in the research market, such as the difficulties of launching experiments and the question of whether experiments are even capable of illuminating much about how to improve schooling.\textsuperscript{21} In our view, demand-side issues may also be important. Specifically, the demand for good evidence on what works in education may be limited because of weak incentives for academic achievement as well as a lack of awareness, on the part of both consumers and producers of such research, about what constitutes high-quality research.

**The Market for Educational Outcomes**

The demand for academic achievement has traditionally been limited for a variety of historical, cultural, and structural reasons. Throughout much of the previous century, for example, schools were expected to educate only the elite to high levels of achievement.\textsuperscript{22} In addition, there is evidence that, in comparison with parents in other societies, parents in the United States place relatively less emphasis on achievement outcomes than on other nonacademic outcomes such as self-esteem and creativity.\textsuperscript{23}
Perhaps most important, however, the organization of schooling in the United States provides little accountability for educators. Most American children receive their elementary and secondary education in public schools. In principle, the country’s public school system may be different from other public sector monopolies, such as the postal service, in that schooling is produced by local rather than state or federal governments, so that families who become dissatisfied with their local schools have the option of “voting with their feet” and moving into new school districts.

In practice, however, most public school administrators have faced relatively little pressure for performance. First, many families are mobility constrained. Even for middle-income families, moving to a different school district catchment area or sending one’s children to private schools are costly exit options. Second, the performance pressures confronting educators from the threat of parent exit may be muted because parents may find it difficult to distinguish between the effects of the school’s “value added” on average test scores and those from the characteristics of a school’s clientele (the student body). This information problem may help explain why most parents indicate that they are satisfied with their children’s school despite achievement test results for public schools as a whole that paint a more worrisome picture. One recent study argues that to the extent to which parents can distinguish between school effects and peer effects, parents may, on average, put more weight on the latter.

If it is true that, at least historically, educators have faced limited pressure for performance, then it should not be surprising that there has been little demand for new and better information about how to improve the technology of schooling. For example, superintendents and principals who face limited accountability for the performance of their students will, in turn, have limited incentives to search for research that indicates which reading program is the most effective and, similarly, be more likely to be swayed by fads in curriculum or pedagogy. Principals who do face pressure to improve their performance will have a strong incentive to change for change’s sake, if only to provide evidence that they are doing something. Change in this case may provide a sufficient signal to parents that the school is responsive to their concerns if parents are themselves unable to determine which policy or programmatic changes are effective.

The Market for Educational Research

In cases in which there has been demand for performance, lack of information about what constitutes high-quality research has contributed to the
undersupply of such research. In this way, the market for education research suffers from a case of what economists refer to as asymmetric information. Research producers must choose among a wide variety of production technologies (research designs), and high-quality research (including quasi experiments as well as randomized field trials) is typically more costly to produce than low-quality research. Education practitioners and their clients are poorly equipped to independently assess the quality of the evidence presented to them on what works in education. The result of such uncertainty about product quality among consumers can lead to a situation in which “lemons” dominate the research market.

Although consumers in the market for medical services often use credentials or accreditation as signals of provider quality, this approach is unlikely to be helpful in education research when so much of the product is generated by credentialed faculty at respectable universities. For example, Charles Murray and Richard Herrnstein’s claims in *The Bell Curve* about heredity, IQ, and government efforts to reduce inequality have been subject to criticism from a number of distinguished academics, including Robert Hauser, Arthur Goldberger, Charles Manski, and James Heckman. Should the public believe a professor from Harvard or tenured faculty from the University of Wisconsin, Northwestern University, and the University of Chicago? Focusing on peer-reviewed publications is unlikely to provide much help to research consumers, given the information problems that even research producers face and the proliferation of peer-reviewed journals.

Unlike the market for evidence on effective medical treatment and devices, the market for evidence on educational interventions receives little regulatory attention from the national government. Currently, the U.S. Food and Drug Administration (FDA) requires that pharmaceuticals and many medical devices be subjected to randomized field trials before they are brought to market. The FDA notes that it regulates the medical research market, in part, because of a medical research literature on the limits of nonexperimental medical research. In contrast, until the recent creation of the What Works Clearinghouse, the Department of Education seems to have devoted relatively little effort to the problem of helping research consumers distinguish between more and less credible findings.

The quasi-governmental National Academy of Sciences and other professional organizations periodically commission independent review panels to synthesize research literatures with important policy applications, which can sometimes have useful effects in guiding educational practice. Such panels are often cumbersome creatures that take many years to organize and implement.
Moreover, the desire of at least the National Academy of Sciences to have different perspectives represented on panels, as well as the encouragement to panels to try to develop consensus, sometimes leads to softened conclusions and thereby reduces the value of the academy’s reports.

We might hope that the media could help fill the role of an independent synthesizer of education policy research that operates in real time. In practice, however, reporters do not seem to be well equipped or empowered to make independent assessments of competing research claims. Too few reporters who cover areas that are informed by social science research actually have much social science training. Another factor is that authorial opinion is currently highly circumscribed for most news reporters, which, in turn, means that news articles are limited to a recitation of competing arguments from those who hold different views. (One friend of ours, frustrated by this media practice in his own area of expertise, once complained to a newspaper reporter, “How come when NASA launches a rocket into space from Cape Canaveral, you don’t feel obligated to call the Flat Earth Society for a rebuttal?”)

Yet the market for education research differs in an important way from other markets that have problems of asymmetric information. In the market for used cars, for example, it is assumed that the seller knows the quality of the car but the buyer does not. As noted earlier in this paper, information problems on the part of consumers is, in itself, enough to lead to an undersupply of high-quality research. In the field of education, there is the additional problem of considerable uncertainty and debate about what constitutes high-quality research, even among the producers of such research.

This uncertainty is apparent on several levels. At the most fundamental level, there is continued debate among education researchers about whether experiments are better than other methods for understanding causal relationships in education. Even among those who accept the idea that randomized experiments are better than the alternatives, however, there is uncertainty about the magnitude of these benefits and how such benefits may vary with the nature of the program being evaluated, the population being studied, or the outcome being examined. It is quite possible, for example, that randomized experiments are much more important for evaluation work in particular circumstances because of the nature of the data available for use in nonexperimental studies or the way in which individuals select or are selected into treatment. An understanding of these factors—the magnitude of the benefits to randomized evaluation and the circumstances under which such research designs are most useful—is crucial because of the costs involved in randomized trials or other rigorous research designs. In the area of education, implementing randomized
trials often requires a great deal of effort, political savvy, and money. Natural experiments require that investigators immerse themselves in the institutional details of how school systems operate in order to find credibly independent variation across students or schools in receipt of the policy or program of interest. In many cases, both random and quasi-random experiments require new data collection.

For this reason, in certain circumstances, it may be more efficient from a social policy perspective to evaluate a program using nonexperimental methods. Robert Slavin makes this point in discussing the efforts to evaluate Success for All, a randomized experimental evaluation in which schools required substantial financial incentives to participate. As Slavin notes, the extra expense associated with evaluating the program through a randomized field trial could have supported a much larger evaluation that used nonexperimentally matched pairs of schools. The extra funding dedicated to running a randomized experiment would most likely reduce bias to some degree, but dedicating these funds instead to expanding the sample size of a nonexperimental evaluation would have benefits in the form of more precise estimates, larger samples that could support subgroup analyses (for example, whether Success for All has differential effects by students’ family income or race and ethnicity), or even a qualitative component to help shed light on the mechanisms through which the program affects student achievement. Understanding the magnitude of the benefits to randomized evaluation (not just knowing that these benefits are positive) is crucial for the decision about how best to evaluate Success for All or any educational program.

Most of the debate about the ability of experiments and other research designs to highlight causal policy effects in education is conducted at the conceptual level. Yet in principle the relative predictive power of different types of research designs could be subject to empirical investigation. Over the past several decades, the fields of medicine, statistics, and economics have generated a growing empirical literature about the conditions under which different types of nonexperimental research methods are able to reproduce experimental findings. The importance of developing a similar literature within education is taken up in the final section of this paper.

The information problem among both producers and consumers in the market for education policy evaluation may be fundamental to the weak demand not just for good education research but also for any education research. If research consumers are unable to assess the quality of evidence, and quality and cost are positively correlated for education evaluation work, then the average quality of education policy research is likely to decline over time. A decline
Recent Policy Changes

Recent federal policies intended to improve the quality of education research target both the supply and demand sides of the research market. Although both strategies are worth pursuing, in the long run efforts to stimulate the demand for higher-quality education research may be more important and promising.

The Education Sciences Reform Act of 2002 seeks to directly stimulate the supply of high-quality education evaluations by changing the U.S. Department of Education’s research activities. The act establishes the new Institute of Education Sciences to “conduct and support scientifically valid research activities, including basic research and applied research, statistics activities, scientifically valid education evaluation, development, and wide dissemination” and to “promote the use, development and application of knowledge gained from scientifically valid research activities.” “Scientifically valid research activities” are defined in part as those that “employ experimental designs using random assignment, when feasible, and other methodologies that allow for the strongest possible causal inferences when random assignment is not feasible.” The two key parameters left undefined are the conditions under which randomized experiments should be preferred to nonexperimental methods and the types of nonexperimental research designs that yield the “strongest possible causal inferences when random assignment is not feasible.” In the absence of any empirical evidence on these two points, almost any research funding decisions could be justified, in principle, under the language of the Education Sciences Reform Act.

Consider, for example, the initial passage in 1965 of the Elementary and Secondary Education Act, which included what was at the time an “unprecedented evaluation requirement.” The language authorizing Title I expenditures states that “effective procedures, including provisions for appropriate objective measurements of education, will be adopted for evaluation at least annually of the effectiveness of the programs in meeting the special educational needs of educationally deprived children.” The meaning of “appropriate objective measurements” was left unclear, and nearly forty years
after passage of the original act there is still little in the way of any rigorous
evaluation of Title I.

In the best-case scenario, the IES will be fully committed to the support of
strong research designs and will navigate the bureaucratic, political, and other
operational challenges that have plagued research operations in the Department
of Education in the past. Even under these conditions, the degree to which the
IES can increase the supply of high-quality education evidence will be limited
by the federal government’s modest expenditures on education research. In
recent years, for example, the research budget of the U.S. Department of Edu-
cation’s Office of Educational Research and Improvement has been on the
order of $130 million. These research expenditures are minuscule in relation
to both total nationwide spending on K–12 schooling, equal to $454 billion in
2001 (which represents 4.5 percent of GDP), and federal spending on med-
ical research, the National Institutes of Health having been authorized by
Congress to spend $27 billion for fiscal year 2003.

Ultimately, what might be more important for education research are recent
federal efforts that directly stimulate the demand for high-quality research and
may thereby affect education research beyond what is funded by the federal
government. Since at least the early 1980s, states have experimented with
accountability reforms in an effort to enhance the incentives for performance
that educators face. Against this backdrop came the No Child Left Behind
Act, which, among other things, seeks to collect and disseminate information
on school performance, including the performance of the subsets of students
within a school who are economically disadvantaged, have disabilities, are
minorities, and have limited English language proficiency. Schools that fail to
meet the performance standards under the NCLB are required to develop
improvement plans and are eligible for technical assistance, and their students
are offered the chance to move to different public schools. In the ideal scenario,
these state and federal policies will increase the incentives for educators to
improve student performance, which, in turn, will provide educators with pow-
erful incentives to obtain more and better information about how to improve
the education “production function.”

How likely is this ideal scenario? One concern is the difficulty of isolating
the causal effects of educators on test scores from the effects of local conditions
outside of their control. Average test scores measured at a particular point in
time will be heavily influenced by student characteristics, which may be a
problem even for test scores that are measured separately for subgroups of
students, as mandated under the NCLB. On the other hand, year-to-year
changes in test scores that help net out the effects of student characteristics are
highly volatile and susceptible to a different set of factors that are also outside the control of school administrators. Other key assumptions behind these accountability reforms include the notion that educators have the resources necessary to improve student performance, that the performance targets are set appropriately (and particularly that they will not be watered down in the face of political pressure), and that the incentives provided to educators are meaningful. Recent empirical work has also documented a variety of perverse responses to accountability reform, such as teacher cheating, increases in the classification of students as disabled, suspension of low-performing students, and even changes in lunchroom menus to boost short-term student performance. These are all large issues that we do not consider here. Suffice it to say that accountability reforms will spur demand for education research only if they successfully push educators to improve performance, which, with current state or federal policies, is not a given.

More certain impacts on the demand for education research come from recent efforts on the part of the federal government to increase the use of evidence in education practice and define what constitutes acceptable evidence (that is, to implicitly regulate the research market in education). The Reading Excellence Act of 1999 provides funding for reading and literacy grants to “improve the reading skills of students, and the instructional practices for current teachers . . . through the use of findings from scientifically based reading research,” defined in part as research that “employs systematic, empirical methods that draw on observation or experiment.” A similar requirement is included as part of the NCLB’s Student Reading Skills Improvement Grants, under which grantees are required to use funds to implement a system “based on scientifically based reading research” (with a definition similar to that in the Reading Excellence Act).

The requirements included in the Reading Excellence Act and the NCLB use overall federal spending for the provision of education (not just funding for education research) as leverage to spur the demand for high-quality evidence. The force of these requirements may ultimately be undermined somewhat by the option for grantees to use methods supported by “observational” research. As Slavin has noted, there is also an important distinction between methods that have been evaluated and shown to be effective and methods that are “based on scientific findings.” For example, a large body of social science research has documented a strong negative correlation between family income and student academic achievement, which under the language of the Reading Excellence Act and the NCLB could be used to justify an intervention that provides extra cash assistance to low-income families with school-age children.
Yet research that directly evaluates the causal effects of increasing the incomes of poor families finds at best only modest improvements in children’s schooling outcomes.\(^57\) One explanation for this apparent paradox is that family income may simply be a proxy for adult characteristics that are associated with both labor market and parenting success.

In any case, despite these concerns, our recent conversations with staff at a number of leading research firms suggest that there is, in fact, increased demand by the producers of curricular materials to have their products evaluated by randomized experiments. These provisions effectively prime demand for education research and then establish a minimum floor for the quality of this research.

**Recommendations**

If the problem with current education policy evaluation is limited supply of high-quality evidence, it is natural to think that the most straightforward solution is to increase federal funding for good research. This is the approach behind the creation of the Institute of Education Sciences, in the U.S. Department of Education, which has the potential to increase the number of randomized experiments used in education research. The impact of this supply-oriented strategy could be enhanced by extending the emphasis on experimentation to other federal agencies that fund education research and, more generally, by expanding total federal funding for education policy research.

There would also be some value, in our view, in reorienting federal funding strategies from a focus on funding streams (such as Title I) to a focus on specific educational programs or policies. In part because evaluation requirements are often included in legislation that provides funding but does not mandate a particular use of the funding, many studies of federal education programs end up examining the effect of additional spending rather than assessing the effect of a particular educational curriculum (for example, a particular reading program) or educational policy (for example, lowering class size by 20 percent). Such studies can, in principle, provide useful information about the overall effectiveness of education spending.\(^58\) However, because schools, districts, and states spend the money in a variety of different ways, such evaluations cannot provide useful information about the effectiveness of any one particular program, which, from the perspective of improving educational practice, is crucial.
A less obvious solution for improving the quality of education policy research is to increase demand. One approach to increasing the demand for high-quality education research is to hold schools accountable for performance, as under the No Child Left Behind Act and other state accountability reforms. However, even with stepped-up demand for education research, consumers in this research market will face difficulty distinguishing between high- and low-quality evidence. Slavin notes, for example, that many schools throughout the country are eager to adopt schoolwide reform programs to improve student achievement but face a bewildering choice of options that have been evaluated with varying degrees of rigor.59

One encouraging development that may help solve this information problem in the market for education research is the IES’s creation of the What Works Clearinghouse (WWC), which seeks to help research consumers make some judgments about the quality of evidence available on a given topic. The Clearinghouse faces a number of significant hurdles in this regard, however: How much weight should be given to a well-executed randomized experiment relative to a nonexperimental study, which may have a weaker research design but is more generalizable to the school or student population of interest? How should the weights assigned to experimental findings vary in response to variation across studies in follow-up sample attrition? How should the weights given to nonexperimental studies vary with respect to research design and the prospect of bias from omitted variables? The hope is that the Clearinghouse will be able to develop some system for addressing these difficult questions and that educators and other research consumers will put greater weight on its research syntheses than on the claims made by individual research studies.

An alternative approach that may leave less room for error is to use federal funding for education as leverage to require that educators pay attention to research and to help them navigate the information problems in the market for research by establishing a minimum standard of evidence. This “quality mandate” is similar to the requirements that the FDA imposes for evidence on the effects of pharmaceuticals or medical devices, although the standards for evidence in recent education laws are not as precisely defined as those imposed by the FDA. Right now, this approach is relevant primarily for reading instruction that relies on federal funding, which accounts for only a small share of total K–12 spending in the United States. At the very least, the Department of Education ought to expand this requirement to cover all subjects and perhaps all pedagogical programs more broadly.60 Moreover, state governments, which account for a large and growing share of education spending, should consider following the federal government’s lead in this regard.
Are these recent changes in federal policy any more likely to succeed than previous efforts to improve the quality of education policy research? One important difference between previous and current efforts is that the latter target both the demand and supply of high-quality education evaluation research. Previous efforts have focused instead on reorganizing, reorienting, or reenergizing the branches of the Department of Education responsible for funding education research. These efforts have floundered in part because the federal government historically does not spend much on education research and because of bureaucratic and political problems that have plagued the different permutations of the Department of Education’s research branch. A more fundamental problem is that efforts to increase the federal government’s supply of high-quality education research can succeed only if federal agencies or their external advisers and peer review teams are able to identify high-quality research. As we have argued, it is not obvious that this is currently the case.

Demand-side efforts to improve education policy research circumvent many of the budgetary, bureaucratic, and political problems that face the Department of Education. However, available information about what constitutes credible education policy research is still a necessary condition for increased demand to translate into increased supply of such research. For example, efforts by the What Works Clearinghouse to synthesize available research will involve decisions about how much weight to put on different experimental and nonexperimental findings. Moreover, Department of Education efforts to follow what might be called “the FDA model” of regulating the market for education research will require better information about the biases of different nonexperimental designs in different education applications to establish minimum quality standards.

Of course, policymakers could, in principle, follow a conservative approach and focus narrowly on randomized experiments. But in cases in which experiments are not possible or have not yet been conducted, dismissing nonexperimental findings altogether leaves educators and policymakers in an information vacuum. Moreover, requiring randomized experiments exclusively may lead to the inefficient allocation of government and private research dollars if certain nonexperimental methods can faithfully reproduce experimental findings in some applications.

A top priority for the IES should be to build a research literature in education about the conditions under which different types of nonexperimental research designs can substitute for randomized experiments, similar to those that are developing in medicine or economics. One way to develop such a literature would be to require every new randomized experiment funded by the
IES to conduct what Robert LaLonde has termed an “evaluation of evaluation methods.” Investigators could themselves issue a request for proposals and ask researchers with different perspectives on what constitutes scientific research in education to put their data where their mouths are, in a sense, and attempt to evaluate the experimental intervention using nonexperimental methods.

For example, those who believe in the power of standard regression techniques could attempt to estimate the impact of the intervention of interest by comparing outcomes for the experimental treatment group with various potential nonexperimental comparison groups. Qualitative investigators could propose to conduct in-depth case studies of a randomly selected set of treatment or control subjects and offer their own assessment of the intervention’s effects, including some discussion about potential heterogeneities in treatment impacts. Slavin suggests that the Department of Education hold a “design competition” for education reform models. What we are suggesting is in some sense a design competition for education evaluation methods, to better understand when, where, and why different education policy research designs are able to yield valid causal inferences about intervention effects. This type of research literature might substantially improve the effectiveness of recent federal efforts to improve the quality of education policy research, as well as begin to develop consensus within the research community about what constitutes credible causal evidence in education policy evaluations.

In sum, we believe that what is new and different about recent policy changes to improve the quality of education policy research is the focus on the demand as well as the supply side of the research market. The ability of research producers to respond to stepped-up demand for more and better education policy research will hinge critically on the ability of everyone involved with the research enterprise to understand what constitutes credible causal evidence under what specific conditions. As policymakers continue to refine their efforts to improve the demand and supply for better education policy evidence, our hope is that they also pay careful attention to solving this fundamental information problem as well.
Comment by Robert Boruch

Brian Jacob and Jens Ludwig have developed a fine paper. I have no major disagreements with what they have said. I do, however, have suggestions that amplify their paper’s virtues and may help to reduce its vulnerabilities.

Randomized Controlled Trials versus Quasi Experiments

Jacob and Ludwig declare that it is important, for scientific reasons, to mount comparative empirical studies to explore whether and when the results of randomized controlled trials (RCTs) differ from the results of nonrandomized trials, also called quasi-experimental designs (QEDs). They argue further that such comparative studies will drive up the demand for higher-quality education research, especially randomized trials. Let me suggest ways to build on their declarations.

I agree with Jacob and Ludwig that the 2003 work of Steven Glazerman, Dan Levy, and David Myers is important. But Glazerman and his colleagues do not consider the directions of biases. They examine only absolute values in the bias. They do not consider scenarios in which bias direction ought to be taken into account and in which the direction of bias may be domain specific. That programs can be made to look harmful when they are merely useless is important to some policy people and certainly to some scientists. That some programs can be made to look as if their effects are positive when their actual effects are negligible is also important to some other policy people and to some scientists.

First, we need to recognize that the direction of statistical biases in QEDs can be important and is likely to be domain specific. For example, D. T. Campbell and R. F. Boruch give plausible scenarios of the different ways in which compensatory education programs can be made to look harmful when, in fact, they are merely useless. An obvious and possible reason for a negative bias in ordinary least squares regression estimates based on observational survey data on some interventions is that adjustment variables (covariates, the right-hand side of the equation, and the like) are measured imperfectly. This, since the 1960s, arguably has resulted in biased estimates of the effects of Head Start, the Comprehensive Employment and Training Act, Youth Employment and Demonstration Projects Act, and other programs when analyses were based on linear models applied to data from nonrandomized trials. A second basic condition is omitted variables, of course.
Think now about the uncertainty and magnitude of bias and how this might be domain specific. For instance, Larry Hedges and Amy Nowell, basing their work on within-country studies, and Aubrey Wang, basing hers on cross-country studies in the Third International Mathematics and Science Study (TIMSS), find that boys’ performance on achievement tests is more variable than girls’.67 These studies are based on probability sample surveys.

We do not know why this holds for math across most of thirty countries in the TIMSS. But this domain specificity, gender being the domain, suggests that our ability to forecast for boys is inferior to that for girls, at least at times. The variability in bias, when an estimator is, in fact, biased, may then also be larger when the target is boys as opposed to girls.

Thomas Fraker and Rebecca Maynard’s comparative study of QEDs versus RCTs is among the few that attends to domain differences.68 Their work invites an exploration of the prospect that the bias in nonrandomized trials may be lower when the target is mainly women (in the Fraker and Maynard study, those who were receiving Aid to Families with Dependent Children) as opposed to mainly young males (in the same study, youth who were involved in Supported Work programs). The complications are obvious. Nonetheless, one implication is that, if we are interested in whether estimators are biased and, if so, by how much, domains may count.

**Coupling Randomized Controlled Trials and Quasi Experiments**

Jacob and Ludwig further suggest that randomized trials might routinely be yoked with nonrandomized studies (quasi experiments) so as to generate evidence on biases. Again, I agree with the idea. Execution, however, is another matter. Their declaration is related to discussions undertaken during the 1990s about embedding RCTs in probability sample surveys and about coupling randomized and nonrandomized trials.69

The idea behind my 1998 study with G. K. Terhanian lay in trying to couple policy research agencies that mount randomized trials, which in 1998 was the U.S. Department of Education’s Planning and Evaluation Service, with agencies, such as the National Center for Education Statistics, that do surveys. This idea may have been prescient—or just bad. The better organizational approach has been taken by Reid Lyon at the National Institute of Child Health and Human Development, Grover “Russ” Whitehurst at the Institute of Education Sciences, and others in building “coalitions of the willing” who seek better evidence by pooling experience and intellectual resources that focus on randomized trials (see Lyon’s paper in this volume). Entities in the ambit of
the Institute of Education Sciences, which includes the National Center for Education Statistics, may be better positioned to yoke surveys and trials in creative ways.

The Demand for Better Evidence and Empirical Comparisons

Jacob and Ludwig argue that doing comparative studies that compare estimates of the effects of interventions based on RCTs against estimates based on QEDs will increase the demand for better evidence—notably, by increasing the pressure to run randomized trials. Perhaps they are right. For some people, however, the most effective illustrations lie in medicine and hinge on identifying badly designed studies on interventions that killed people. Early examples led to the Estes Kefauver hearings, the U.S. Food and Drug Administration rules on RCTs, and the sequelae. The Coalition for Evidence-Based Policy and the Institute of Education Sciences give other examples, to good effect, from the twenty-first century, including references to the randomized and nonrandomized trials on hormone replacement therapy to inform the public about this topic.

Misleading evidence does not often involve immediate death, as in the medical examples. Slower deaths are marked in different ways. Serious examples of misleading evidence need to be identified in the education, mental health, and crime sectors. Indeed, the Campbell Collaboration’s Fifth Annual Colloquium in Lisbon (in 2005) will include a session on the topic, organized by Anthony Petrosino. Putting aside the dramatically morbid examples in medicine, health researchers continue to produce interesting work that compares the results of randomized trials against results of nonrandomized trials. Regina Kunz and A. D. Oxman, whom Jacob and Ludwig cite, are a case in point.

Access to Information and Micro Records

How accessible is information for comparative empirical studies of the sort that Jacob and Ludwig, and other contributors to this volume, have proposed? Glazerman, Levy, and Myers, Valerie Nelson and C. F. Turner, and others had access to micro records. W. R. Shadish, J. K. Luellen, and M. H. Clark managed to mount prospective randomized and nonrandomized trials in parallel; consequently, they had not only micro records but also an intimate understanding of the isomorphism in their comparative study. Nelson and Turner’s review of reanalyses of micro records from studies of programs under the Youth Employment and Demonstration Projects Act programs went a long
way toward informing the U.S. Department of Labor that nonrandomized studies, as designed in the 1970s, did not produce estimates that come close to those from randomized trials. A similar sequence might reinforce public interest in randomized trials in education.

In contrast, Kunz and Oxman had access only to reports on randomized and nonrandomized trials identified by the Cochrane Collaboration in health care. They did not have access to micro records from the studies. Their review depended on reports that gave only statistical summaries of original analyses.

The Institute of Education Sciences’ What Works Clearinghouse is developing information at both the study level and the multistudy level that could be useful in comparative empirical research of the kind that Jacob and Ludwig suggest. For instance, all impact studies conducted in the United States on middle school mathematics that meet a minimum standard for causal inference will be identified as RCTs or QEDs. As of 2004, peer-assisted learning and at least four other topics will be similarly identified. The What Works Clearinghouse, as currently configured, will identify and describe the relevant studies and assess the dependability of the evidence that is provided in published (peer-reviewed) reports and unpublished reports on studies of the effects of interventions.

The What Works Clearinghouse will, in effect, provide some material to Jacob and Ludwig and to anyone else who cares to do research on randomized trials versus quasi experiments. It will also provide information that will help analysts determine whether they want to obtain micro records generated in each study.

**Federal Agency Support**

Jacob and Ludwig encourage the Institute of Education Sciences to support methodological studies that compare results of randomized and nonrandomized trials. Given that the institute’s policy demarcates RCTs and QEDs, in the What Works Clearinghouse and elsewhere, it is reasonable to assume that it understands the encouragement, regardless of whether it chooses to invest in such work. In fact, such work might be supported under its unsolicited grants program. More important, perhaps, we need to recognize that the Institute of Education Sciences is only one of a number of federal agencies that support, or could support, methodological research of this kind.

Some of the institutes of the National Institutes of Health can support such work, for example. In the AIDS arena, for instance, federal staff and others were interested in the 1990s in whether QEDs’ estimates of the effects of AIDS
education programs could match estimates based on RCTs. Some colleagues at the National Institute of Mental Health, the National Institute of Allergy and Infectious Diseases, and elsewhere are still interested. And they should be, because RCTs are hard to do, and our capacity to create plausible model-based alternatives to randomized trials is substantial. Federal research agencies in the United States can learn to cooperate on better methodological studies on this topic, as they have learned to cooperate in research on other important topics.

Quality of Important Features of the Study

The quality of the study design is important, but it does not address important measures of the study’s functionality. Does the evidence at hand support causal conclusions? To be sure, a well-designed randomized trial allows causal inferences, other things being equal, that a quasi experiment rarely permits. However, a high-quality RCT design that is poorly executed may produce evidence that is equivocal, and a nonrandomized trial that has been designed and executed well can produce evidence that may be equivocal but is nonetheless usable. The latter may merely suffice—rather than satisfy, as would evidence of a higher quality. Comparative empirical studies may help us learn more about this.

The What Works Clearinghouse has had to focus on functionality. It attends to whether the evidence is sufficient to support causal inferences. The quality judgments depend on both the study’s design (the Clearinghouse regards RCTs as a priority) and its execution, the analysis of resulting evidence, and other dimensions apart from design.

Regression Discontinuity Designs

J. D. Angrist and Victor Lavy have studied the effect of class-size reduction in Israeli schools.\textsuperscript{80} The approach, invented by Donald Campbell and perhaps also Arthur Goldberger, which was called “regression discontinuity” during the 1970s and 1980s in the literature cited by Angrist and Lavy and “point displacement” in later publications, is important. The regression discontinuity approach results in unbiased estimates of the effects of an intervention if the intervention’s deployment is based strictly on a continuous and reliable measure of need or merit for the intervention (that is, a sharp cut point) and the empirical relation between the continuous measure of need or merit and the continuous outcome measure in the absence of the intervention is known.
Most users of regression discontinuity assume or know that the relation between the selection variable ($x$) and the outcome variable ($y$) is linear. Having evidence from earlier observational studies on the linearity is good. If there is no earlier evidence on linearity, then there may be problems. For instance, in the 1970s, James DeGracie (of the Mesa, Arizona, school district) and I tried to estimate the effects of Title I–related programs in Mesa, based on regression discontinuity designs. We discovered that the relationship between the selection variable and the outcome variable was not clearly linear from one grade level to the next. At the time, we speculated that ceilings and floors on the tests induced nonlinearity in the empirical relations. We lacked the resources to dig into nonlinearity in relationships in test scores from the beginning of one semester to the next. Nor did we have the time and resources to publish our results. The publication bias is as severe here as in other research contexts.

A strict cut point is important in a regression discontinuity design. But some administrators do not like strict rules. They want wiggle room, and they take advantage of it. So fuzzy cuts become important. Noncompliance in this design is likely to be at least as difficult a problem as noncompliance in RCTs.

Capacity Building and Randomized Controlled Trials

Jacob and Ludwig assume that randomized controlled trials in the education arena can be well designed and executed. The assumption is not always justified, at least to judge from proposals submitted to the William T. Grant Foundation. The board of trustees’ members (of which I am one) and the foundation’s senior executives and advisers frequently confront proposals for RCTs that routinely invoke hierarchical linear models and instrumental variables. In the past, invocations were based on studies using Restricted Maximum Likelihood Factor Analysis (RMLFA) and Linear Structural Relationships (LISREL) software. In the interim, the supplications referred to propensity scoring and selection models.

These invocations to interesting, industrious, and creative statistical analyses were often not well informed, though many of the proposal writers are well positioned to cooperate with school administrators, researchers, and policymakers producing evidence about what works. The William T. Grant Foundation decided to invest in capacity building at the technical level. In particular, the foundation arranged sessions in 2004 on hierarchical linear models in the context of cluster (place) randomized trials at meetings of the Association of Policy
Analysis and Management and the Society for Research in Child Development and organized a specialized workshop at the University of Michigan. The origins of the foundation’s effort lie partly in related Campbell Collaboration efforts undertaken with support from the Rockefeller Foundation.

The U.S. Department of Education’s Institute of Education Sciences has also initiated capacity-building efforts, investing dollars that far exceed the resources of a private foundation. Graduate students will benefit from learning about how to generate better evidence, as will master teachers at schools. In the long term, children will benefit.

A Little Homage to Historians

Some historians have contributed seriously, in ways that I admire, to our thinking about evidence in education. They include Diane Ravitch, who has organized the Brookings Institution’s annual conference on education policy, and Maris Vinovskis, Patricia Graham, and Ellen Condliffe Lagemann, among others, who have written fine histories of education and education policy and contributed to the federal government through their services in agencies or through advisory boards of the National Academy.

More to the point, as a closet historian I honor colleagues with the earliest reference I can find to comparative empirical studies on whether and how the results of RCTs and QEDs differ. The little gift is F. Yates and D. A. Boyd’s agricultural research. Estimates of the effect of spreading farmyard manure on potato crop yield were large and positive when based on randomized trial. The estimates based on passive survey data and ordinary least squares regressions were large and negative. The difference is possibly related to differential skills of farmers in raising animals and spreading manure—a variable that was omitted in the regressions.

Conclusion

Jacob and Ludwig’s suggestions are helpful. If they are acted on, taking into account the further suggestions I present here, more can be learned about why randomized controlled trials are a dependable source of evidence about the effects of interventions and when nonrandomized trial results might approximate the results of randomized trials or when the two can be productively coupled for other reasons. This learning is important for science, for policy, and for the well-being of children.
Comment by Maris A. Vinovskis

Brian Jacob and Jens Ludwig ask an important question in their timely and thoughtful essay: Will recent federal policy changes succeed in improving the quality of education policy research? To address this issue, they consider the impact of the No Child Left Behind Act and the creation of the Institute of Education Sciences (IES). They believe that the IES is improving research and policy evaluations by funding randomized field trials and creating the new What Works Clearinghouse. Yet they argue that the work of the IES, by itself, will not be enough to provide a sufficient supply of high-quality education research.

Jacob and Ludwig call for stimulating demand for high-quality research by fostering public and policymaker interest in improving student achievement. With the No Child Left Behind Act holding local schools and educators accountable for improving student achievement, policymakers and educators may become more interested in adopting education practices and programs that have been scientifically proved effective. Given the high cost of randomized field trials and the limited research dollars available, Jacob and Ludwig also advocate finding less expensive, but still highly effective, ways to pursue high-quality research and program evaluations.

They properly appreciate the recent efforts to generate high-quality education research by passage of the No Child Left Behind legislation and the establishment of the Institute of Education Sciences. They are also to be commended for supporting the use of more randomized field trials in education research as well as calling for the development of less expensive, but no less effective, alternative analytical tools. Perhaps some of the current federal education research and evaluation strategies and practices, however, should be explored even further—especially in light of earlier, disappointing federal efforts to provide high-quality research, program evaluations, and systematic development.

What Kinds of Policy Research and Evaluation Are Needed?

Many different types of research and evaluation projects might be pursued. The nature and scope of these studies will vary considerably depending on whether one focuses narrowly on a particular intervention or includes broader education policy issues. Jacob and Ludwig provide several examples of evaluations of such issues as the benefits of smaller class size, whether a particular
reading curriculum leads to improvements in reading comprehension, and whether “pull-out” programs are more effective than “push-in” programs for students with learning disabilities. These evaluations are particularly amenable to randomized field trials or perhaps other less expensive analytic strategies.

But what about analyzing more comprehensive federal education initiatives, such as systemic reform, Goals 2000, or No Child Left Behind? What kind of evaluation strategies and analyses will be needed for these programs? Although randomized field trials can be helpful for some of the smaller, individual components of those packages, overall strategies for evaluating broader and more comprehensive initiatives are even more challenging—especially given the limited funds provided for evaluation and the pressure from policymakers to produce at least preliminary results within a few years.

Who, within the federal government, is funding systematic analyses of broad alternative policy approaches today? In the past, some of the work was done at research and development centers. During the first Bush administration, for example, key ideas about systemic reform were sketched out by Marshall “Mike” Smith and Jennifer O’Day as part of their federally funded work at the Policy Center of the Consortium for Policy Research in Education. With the diminished size and changing scope of research and development centers today, will the IES support comparable policy explorations through some of those grantees? The congressionally mandated independent review panel for the Title I program is working with the current administration to evaluate the No Child Left Behind Act, but it is not analyzing broader, alternative education policies. Should rigorous consideration of broader policy issues be supported by the IES, or is this an area of policy analysis better left to funding from nongovernmental sources? Certainly the executive branch and the Congress will consider larger policy issues during the reauthorization of the No Child Left Behind Act in a few years; but who should be helping to provide the lawmakers with the in-depth background information and policy analyses they will need? Much of the IES’s new emphasis is funding several useful multiyear randomized control trials in areas such as reading comprehension research and preschool curriculum evaluation; but the agency does not appear to have plans for long-term developmental projects that take promising practices, subject them to intensive experimental testing, and evaluate their effectiveness in different settings and under diverse circumstances.

Interestingly, the current Bush administration’s most publicized and successful endeavor, their early reading programs, grew out of a thirty-year investment in reading by the National Institute of Child Health and Human Development (NICHD), ably led for many years by Reid Lyon. Rather than
being based upon a few short-term, randomized field trials, the NICHD reading studies represent a multimillion-dollar annual research and development program that has spanned several different administrations. Yet a comprehensive strategy for program development does not appear to be the template for most IES initiatives at this time.

In terms of conceptualizing more rigorous, systematic development, the 108th Congress and the current administration seem to have forgotten the valuable recommendations made more than three decades ago. In 1973 the Brookings Panel on Social Experimentation hosted a conference exploring the ongoing planned variation analyses of Project Head Start and Follow Through. Based upon the conference deliberations, the organizers, Alice Rivlin and Michael Timpane, called for a strategy of education development that could be subdivided into five stages and would require ten to twelve years to complete:

The strong consensus of the conferees that planned variation experiments of the future should have clearer objectives, more time, and cleaner statistical design was given concrete expression in a proposal for a five-stage process of development and testing of educational interventions. The experiment would begin as a highly controlled investigation at a single site involving random assignment to control and treatment groups and careful observations of inputs and outcomes. If the intervention appeared to have appreciable positive effects under these conditions, a couple of years would then be devoted to developing it further, creating a training program for teachers and instruments for measuring the program’s implementation and outcomes. The intervention would next be tried out under natural conditions in a small number of sites, close enough to the sponsor’s home base to be supervised without great travel and communication costs, and curriculum, training procedures, and measuring instruments would be revised in light of this experience. Not until after all of this development, small-scale testing, and revision had been successfully completed would a large-scale field test be undertaken to find out how the intervention works under a variety of conditions and with a variety of populations. In the final stage, full results of the field testing and training would be disseminated to those who wanted to adopt the intervention in their own school. The whole cycle, from initial experiment to dissemination of the field-test experience, would take ten to twelve years.

The promising research and evaluation strategy recommended by the 1973 Brookings Panel on Social Experimentation should be rediscovered today. Indeed, the idea of development used to be a key component of federal education research strategy. Many of the research and development centers and the regional education laboratories stressed program and product development—though none of them ever pursued that approach as rigorously and systematically as suggested by the Brookings panel. Yet even the 1994 reauthorization of the Office of Educational Research and Improvement (OERI)—the predecessor agency to the IES—specifically called for systematic
program development. Unfortunately, OERI did not try to implement that directive.\textsuperscript{88} Today neither the 108th Congress nor the Institute of Education Sciences has articulated or implemented a comparable vision of systematic development.

\textit{Fragmentation of Education Research and Evaluation}

One of the recurrent complaints about federal education research and program evaluation is that most of the efforts have been small, short-term projects. Moreover, the larger, long-term goals of the federal agencies have frequently been revised as new administrators arrive; and the Congress continues to mandate numerous small, underfunded studies. The creation of the Institute of Education Sciences is certainly a step in the right direction; but many key federal education initiatives are handled outside the IES and may not be part of any systemic, rigorous undertaking.

The Office of Innovative Improvements, for example, handles a wide variety of important issues ranging from school choice to the teaching of American history. Is the research and demonstration work supported at the Office of Innovative Improvements expected to produce the high quality of scientific research and evaluation mandated by the No Child Left Behind legislation, or do the standards for work produced by the IES differ from those for other Department of Education agencies?

For example, the No Child Left Behind legislation emphasizes certain projects, such as improving early reading programs. The Institute of Education Sciences and the NICHD continue to energetically support that endeavor with substantial research and evaluation. But what about other subjects, such as the teaching of American history? Neither elementary and secondary education students nor adults have the knowledge of our past deemed necessary for good citizenship. Therefore, should there also be rigorous investigation exploring the best ways of teaching American history in K–12 schools?

Thanks to the initiative of Senator Robert Byrd (D-W.V.), Congress is allocating about $120 million dollars annually for improving the teaching of American history. The legislation calls for local schools, in partnership with others, to develop these programs. Although the Office of Innovative Improvements, with IES assistance, has tried to encourage rigorous evaluations of some of these programs, the results to date have been disappointing. Although the American history programs have stimulated interest in the teaching of history and provided much-needed financial support to local school districts, none of the American history projects appears to be utilizing the same rigorous,
long-term development strategies characteristic of the NICHD’s early reading program. Instead, several hundred small, three-year projects are being funded that provide untested professional training for small groups of K–12 social studies and history teachers.

In other words, though the No Child Left Behind Act repeatedly calls for rigorous and scientifically based studies of the curriculum and teaching practices in areas such as reading, it seems that a different set of standards and expectations is applied in other subject areas, such as American history. Has the administrative decision to locate the oversight of a particular subject area in the Office of Innovative Improvements affected the nature and scope of work undertaken? Although there are certainly not enough resources to research every issue as rigorously as the development of the reading programs, are we making the most effective, long-term use of the approximately half-billion dollars that has been spent on learning how to improve the teaching of American history?

**Monitoring the Quality of Federal Education Research**

Despite periodic efforts to improve governmental oversight, earlier federal research agencies such as the National Institute of Education and the Office of Educational Research and Improvement did not do an adequate job of monitoring the quality of education research and evaluation. The Institute of Education Sciences has taken an important step forward by creating the What Works Clearinghouse. The Clearinghouse will rigorously and systematically evaluate education research studies in selected areas, but it will not monitor or assess the quality of most of the ongoing research of either the IES or other Department of Education agencies.89

In 1994 Congress, concerned about the quality of education research, mandated that the OERI, in conjunction with its National Educational Research Policy and Priorities Board, establish standards for the conduct and evaluation of research. The OERI was required to institute those standards in three phases: guidelines for evaluating applicants, procedures for exemplary and promising practices, and standards for periodic evaluation of the performance of all OERI-funded activities.90

Although the OERI did a reasonably good job of addressing the first and second phases of the mandated standards, it paid much less attention to creating and implementing standards for evaluating the research and evaluation work funded by the agency. For example, the work of the regional education labora-
tories and the research and development centers was not rigorously examined; nor were the third-year or final reviews of those institutions well done.\textsuperscript{91}

What is the IES doing to evaluate the quality of the research and evaluations it continues to fund? Some individual projects have been assessed, but apparently no overall standards, comparable to the earlier phase-three mandates, have been promulgated.\textsuperscript{92} Moreover, although the 106th and 107th Congresses had criticized the earlier evaluations of the regional education laboratories, the IES has not undertaken a third-year assessment of those institutions—in part because the Bush administration has been unsuccessful in abolishing those institutions and in part because monies were not set aside for an evaluation. The regional educational laboratories currently receive $66.7 million annually, about 15 percent of the entire IES budget (and equivalent to 40 percent of the IES’s entire basic research, development, and dissemination budget). As the next administration and the 109th Congress plan the future of the regional educational laboratories, will they have up-to-date, reliable information on the quality of the work produced by those institutions?

As the first four years of the Bush administration draw to a close, there still is no comprehensive set of standards or plans for evaluating the quality of the work produced by the IES or other agencies such as the Office of Innovative Improvements. Thus though important progress has been made in some areas, such as the creation of the What Works Clearinghouse, much remains to be done in setting and implementing high-quality standards for federally funded research projects not scheduled to be reviewed by the Clearinghouse.

\textit{Politics and Education Research}

During the past four decades, federal education research has frequently been hampered by political controversies. For example, the Council for Educational Development and Research, with the assistance of key members of Congress, forced the National Institute of Education and the OERI to fund regional education laboratories, even though the quality and usefulness of their work was questionable. The Reagan administration tried to purge from the National Institute of Education researchers and staff who they thought were too biased toward a liberal agenda. During the Clinton administration, the OERI occasionally engaged in repressing or delaying the timely release of evaluations that might have adversely affected the interests of the regional laboratories.\textsuperscript{93}

Should we be concerned that politics may continue to intrude inappropriately into the operation of the Institute of Education Sciences? Many observers
have called for a completely independent federal education research and evaluation agency. The Education Sciences Reform Act of 2002 placed the IES within the Department of Education, but it did provide for a six-year term for the presidentially appointed director, Grover “Russ” Whitehurst.94

Although the IES under Whitehurst has received little criticism in terms of its provision of objective, nonpartisan research and evaluations, the situation might be quite different under some other appointee in the future. For example, when key issues, such as the effectiveness of voucher programs, are evaluated and debated, is there a possibility that a future George W. Bush or John Kerry White House might be tempted to directly or indirectly influence the assessment process?

Everyone seems to assume that the findings from rigorous, scientifically based studies, utilizing random field trials, will be accepted by lawmakers. But might politics sometimes still trump rigorous research and evaluation? For example, the Even Start Program, which fosters the literacy of young disadvantaged children and their parents, was evaluated by Abt Associates using, whenever possible, a random assignment of families to the Even Start project or a control group. The 1995 results from that evaluation were disappointing; though the children and parents in Even Start improved their literacy, those in the control group did equally well.

Even Start was created at the initiative of Representative William Goodling (R-Pa.) and has enjoyed strong, bipartisan support in Congress. Goodling, a well-respected and effective proponent of high-quality education research and evaluation, faced a problem in seeking additional Even Start funding once the disappointing Abt evaluation results were released. At subsequent congressional hearings on Even Start, Goodling and his Democratic allies simply ignored the negative findings from the Abt evaluation and relied upon anecdotal testimony from several satisfied Even Start staff and clients.95 Thus as we applaud the increased employment of random field trials and other rigorous research approaches, perhaps we also need to pay some attention to how the results of such studies might be received and used by educators and policymakers who have a strong personal interest in the outcomes of those evaluations.

Notes


3. Critics of education research often argue that the relative lack of emphasis on program evaluation research within education is itself part of the problem.

4. Our thesis is based on the assumption that there is at least some merit to concerns expressed by the critics of education research. One of the points developed here is that this is itself an issue that is amenable to empirical study and verification. We present some evidence to support this claim in the following section but do not provide new empirical analysis on this point.


8. See Robert Boruch, Dorothy de Moya, and Brooke Snyder, “The Importance of Randomized Field Trials in Education and Related Areas,” in *Evidence Matters*, edited by Mosteller and Boruch, pp. 64–65. As another example, a search conducted on June 5, 2004, of the Campbell Collaboration’s library of randomized or possibly randomized trials under the keyword “education” yielded a total of 1,459 entries (campbellcollaboration.org). This figure is very small in relation to all of the education research articles published by scholarly journals during the past several decades.


14. By collecting rich data on a variety of potential mediators, the best randomized experiments can eliminate potential mediators or provide suggestive evidence on potential mechanisms; see, for example, Jeffrey R. Kling and Jeffrey B. Liebman, “Experimental Analysis of Neighborhood Effects on Youth,” Working Paper 483 (Princeton University Industrial Relations Section, 2004); or Jeffrey R. Kling, Jens Ludwig, and Lawrence F. Katz, “Neighborhood Effects on Crime for Female and Male Youth: Evidence from a Randomized Housing Voucher Experiment,” Working Paper 10777 (Cambridge, Mass.: National Bureau of Economic Research, 2004).


21. For a review of these arguments, see Cook, “Randomized Experiments in Educational Policy Research,” in *Evidence Matters*, edited by Mosteller and Boruch.


24. In the short run, even private school attendance by local students is unlikely to provide much incentive to public school administrators to improve the quality of local schools because the parents of such children will continue to pay property taxes.


26. Jesse Rothstein, “Good Principals or Good Peers: Parental Valuation of School Characteristics, Tiebout Equilibrium, and the Incentive Effects of Competition among Jurisdictions,” working paper (Princeton, N.J.: Princeton University, Department of Economics, 2003). Other studies provide similar evidence on parental preferences. For example, several recent studies suggest that house prices are more strongly correlated with average school test scores, which reflect both peer and school effects, than with within-cohort changes in test scores, which is typically thought to provide a better measure of a school’s value added; see David N. Figlio and Maurice E. Lucas, “What’s in a Grade? School Report Cards and House Prices,” working paper (University of Florida, Department of Economics, 2002); and Thomas J. Kane, Douglas O. Staiger, and Gavin Samms, “School Accountability Ratings and Housing Values,” *Brookings-Wharton Papers on Urban Affairs*, edited by William Gale and Janet Rothenberg Pack (Brookings, 2003), pp. 83–138.
27. See George Akerlof, “The Market for Lemons: Quality Uncertainty and the Market Mechanism,” *Quarterly Journal of Economics* 84, no. 3 (1970): 488–500. For example, suppose a consumer is willing to pay $10 for a low-quality study and $100 for a high-quality study, and that it costs $5 to produce a low-quality study and $90 to produce a high-quality study. If 5 percent of the studies offered initially are of high quality, then apparently the most a consumer would be willing to pay for a study is $14.50. However, for the price of $14.50, producers are willing to sell only low-quality studies (because it costs them $90 to produce a high-quality study). This dynamic will lead to a situation in which only low-quality studies are offered, even though the price consumers are willing to pay for high-quality studies exceeds the cost of producing them. More generally, the equilibrium quantity of high-quality research offered by producers will be determined by the costs of producing low- versus high-quality research and the value of each product to consumers.


30. The FDA requires that the safety and efficacy of drugs be demonstrated by evidence from randomized clinical trials except in cases of serious or fatal diseases for which there are no treatment alternatives other than the (unapproved) investigational medication. As the FDA notes, “Clinical testing isn’t the only way to discover what effect drugs have on people. Unplanned but alert observation and careful scrutiny of experience can often suggest drug effects and lead to more formal study. But such observations are usually not reliable enough to serve as the basis for important, scientifically valid conclusions. Controlled clinical trials, in which results observed in patients getting the drug are compared to the results in similar patients receiving a different treatment, are the best way science has come up with to determine what a new drug really does. That’s why controlled clinical trials are the only legal basis for the FDA to conclude that a new drug has shown ‘substantial evidence of effectiveness, as well as confirmation of relative safety in terms of the risk-to-benefit ratio for the disease that is to be treated.’” U.S. Food and Drug Administration, *From Test Tube to Patient: Improving Health through Human Drugs* (Center for Drug Evaluation and Research, September 1999), p. 18.


32. Slavin, “Evidence-Based Education Policies.”

33. This problem is highlighted by the recent curriculum changes being considered at Columbia’s journalism school; see, for example, Karen W. Arenson, “Columbia Names New Dean for Its Journalism School,” *New York Times*, April 16, 2003.

perhaps the Wall Street Journal, an extremely accurate newspaper, points the way out of this morass. The Journal relies on confidential sources, of course, but it also allows its reporters more latitude in asserting what they know to be true. The paper encourages reporters to incorporate analysis into their reporting rather than compose a stenographic procession of facts, quotations, and official denials, allowing it in many cases to get closer to the truth than its rivals.”

35. See, for example, Cook, “Randomized Experiments in Educational Policy Research.”

36. Slavin, “Evidence-Based Education Policies.”

37. See Cook, “Randomized Experiments in Educational Policy Research,” for a review of these arguments. As Robert Boruch and his colleagues note, “The mathematical conditions (assumptions) under which [a randomized field trial] and a nonrandomized quasi experiment will differ are reasonably well understood. Not so clear are the empirical conditions under which this difference occurs.” Boruch, de Moya, and Snyder, “The Importance of Randomized Field Trials,” p. 70.


40. As Robert Slavin notes, “Policymakers have rarely seen the relevance of research to the decisions they have to make and therefore have provided minimal funding for research. This has led to a declining spiral, as inadequate investments in research lead to a dearth of the kind of large-scale definitive research that policymakers would feel to be valuable, making these policymakers unwilling to invest in large-scale, definitive research.” Slavin, “Evidence-Based Education Policies,” p. 17.


43. Quoted ibid., p. 18.

47. National Institutes of Health homepage (www.nih.gov/about [June 9, 2004]).
48. See, for example, Holding Schools Accountable: Performance-Based Reform in Education, edited by Helen F. Ladd (Brookings, 1996).
56. See Slavin, “Evidence-Based Education Policies.”
58. In practice, because the allocation mechanism for the funding is not random, it is difficult to isolate the causal impact of additional educational resources in this scenario.
59. Slavin, “Evidence-Based Education Policies.”
60. We recognize that defining the exact scope of spending programs subject to this mandate could be difficult. Clearly, there is some funding that could not and should not be subject to this type of requirement, such as funding used to purchase office equipment.
62. See Boruch, de Moya, and Brooke Snyder, “The Importance of Randomized Field Trials.”
63. See LaLonde, “Evaluating the Econometric Evaluations of Training Programs with Experimental Data.”


77. Kunz and Oxman, “The Unpredictability Paradox.”

78. cochrane.org (September 29, 2004).

79. The website is at whatworks.ed.gov (September 29, 2004).


81. The website is at wtgrantfdn.org (September 29, 2004).

82. The website is at campbellcollaboration.org (September 29, 2004).


86. For an introduction to the NICHD reading program, see the paper by G. Reid Lyon, Sally E. Shaywitz, Bennett A. Shaywitz, and Vinita Chhabra in this volume.


93. Vinovskis, “Missing in Practice?”
