THE VALUE OF EXPERIMENTS IN EDUCATION

Grover J. Whitehurst
Brookings Institution
1775 Massachusetts Avenue NW
Washington, DC 20036
gwhitehurst@brookings.edu

Abstract

One of the major story lines of the growth of civilization is the advance of the experiment. From the food we eat to the diseases we conquer to our understanding of how we think and behave, we have profited enormously from an approach that marries our models of the world with tests of their validity through systematic variation to determine cause and effect. These same tools of thought and action are no less critical to the advance of education than to medicine, agriculture, psychology, or transportation. There has been impressive growth in the use of experimental methods in education over the past decade. As a result, much more is known today about what works and what does not. Future progress will be enhanced by research that better explicates process: linking trials to administrative data; serving the interest of local and state officials in examining the impact of their policies and practices; better preservice training for education practitioners in the logic and value of rigorous research; and more generous federal funding for the research enterprise.

1. INTRODUCTION

Prior to the Johnson administration, the federal role in education was small in concept and reality. Although an office of education was established in the executive branch in 1867, its functions were limited to gathering information on the status of education. For the next one hundred years, accretions of the federal role were generally limited to postsecondary education, with the most notable programs being the establishment of land-grant colleges at the advent of the twentieth century, the passage of the GI Bill following World War II, and the creation of the National Defense Education Act following the launch of Sputnik to fund student loans and graduate study. Federal involvement in primary and secondary education began in earnest in the context of the civil rights movement of the 1960s and 1970s with the passage of Title VI of the Civil Rights Act of 1964, Title IX of the Education Amendments of 1972, Section 504 of the Rehabilitation Act of 1973, and the Elementary and Secondary Education Act of 1965 (ESEA). The ESEA remains to this day the principal mechanism through which the federal government is involved in K–12 education, with the most recent reauthorization of ESEA being the No Child Left Behind Act of 2001 (NCLB).

Despite the federal legislative onslaught that began in the 1960s and President Lyndon Johnson's desire to be an education president, the federal touch on elementary and secondary education until recently remained light. Education did not achieve cabinet-level status until 1980, and the greater federal role that cabinet rank symbolized was controversial. Presidential candidate Ronald Reagan campaigned in 1980 on a promise to eliminate the Department of Education, and a similar campaign promise was made by presidential candidate Bob Dole in 1996.

The 2001 reauthorization of ESEA (or NCLB) was a sea change in federal involvement in education. The 1994 authorization of ESEA (the Improving America's Schools Act) emphasized standards and assessments, whereas NCLB shifted the focus to holding schools and school districts accountable for annual achievement targets for student populations identified by minority, disability, and language status via statewide tests. This top-down federal role was new in the history of the nation—one could argue that George W. Bush provided the muscle to turn Johnson's aspirations into reality.

I was serving as U.S. assistant secretary of education for research and improvement when NCLB became law and shortly thereafter was appointed as the first director of the Institute of Education Sciences within the U.S. Department of Education. I was surprised at my popularity among practitioners at the initial meetings the department organized to roll out NCLB. A chief state school officer or a district superintendent would pull me aside and ask what scientific research had to say about professional development, or effective

mathematics curriculum, or . . . (fill in the blank). "I'm desperate to meet the student proficiency requirements of NCLB. Just tell me what the research says works and I'll do it" was the gist of their position.

For most such requests I had to say that I didn't know of any decent research that would be helpful. The comeback from the practitioner, sometimes articulated and sometimes implied, was to question how NCLB could require the use of scientific research findings I was saying did not exist. I'm sure that a number of my interlocutors wondered whether I knew what I was talking about or had set an unreasonably high bar for what constitutes useful research on their questions of interest. Determining whether they were right requires some context.

2. A BRIEF HISTORY OF EDUCATION RESEARCH AT THE FEDERAL LEVEL

In 1971 the President's Commission on School Finance (1971, p. 5) commissioned the RAND Corporation to review research on what was known about what works in education, reasoning that "the wise expenditure of public funds for education . . . must be based on a knowledge of which investments produce results, and which do not." RAND concluded that

> the body of educational research now available leaves much to be desired, at least by comparison with the level of understanding that has been achieved in numerous other fields. (p. 157)

> Research has found nothing that consistently and unambiguously makes a difference in student outcomes. (Averch et al. 1972, p. 154)

In other words, forty years ago there was no evidence that anything worked in education. It was not until the late 1950s that the National Science Foundation and the Office of Education within the Department of Health, Education, and Welfare began to fund education research, so perhaps the dearth of evidence when RAND did its report in the early 1970s should not have been surprising.

As a response in part to the work of the President's Commission on School Finance, Congress created the National Institute of Education (NIE) in 1972 in the Department of Health, Education, and Welfare to provide a credible federal research effort in education. NIE was moved to the Office of Educational Research and Improvement (OERI) within the Department of Education when that department came into being in 1980. A 1985 reorganization of OERI abolished NIE.

Federal investments in education research, while always minuscule compared with investments in research in fields such as health care and agriculture, grew substantially with the founding of NIE and had amounted to more than $2.6 billion through NIE and OERI by the close of the twentieth century (Vinovskis 2001). One would imagine that the creation of a federal education research agency and the increased levels of federal investment would have improved the status and yield of education research by the end of the century. However, 1999 saw the issuance of a report on education research by the National Academies of Science that came to essentially the same conclusions as the RAND report of twenty-seven years earlier:

> One striking fact is that the complex world of education—unlike defense, health care, or industrial production—does not rest on a strong research base. In no other field are personal experience and ideology so frequently relied on to make policy choices, and in no other field is the research base so inadequate and little used. (National Research Council 1999, p. 1)

Why was there so little to show for more than forty years of federal involvement in education research? One possibility is that NIE and OERI were organizationally weak or funded the wrong types of research, or both. In a historical analysis of federal education research, Andrew Rudalevige (2008, p. 18) cites James March's description of NIE as an organization that "came to be indecisive, incompetent, and disorganized" (March 1978). Rudalevige adds the statement of an assistant secretary for OERI, Diane Ravitch, that her "agency itself bears a measure of blame for the low status accorded federal educational research." He caps his point with a quotation from Gerald Sroufe, director of government relations at the American Educational Research Association, that toward the end of OERI's life congressional observers were describing it in "language . . . [that] cannot be printed in a family-oriented academic journal."

Congress acted on its growing frustration with federal management of education research by passing the Education Sciences Reform Act of 2002, which abolished OERI and replaced it with the Institute of Education Sciences (IES). IES was given a greater degree of independence from the Department of Education's political leadership than had been afforded to OERI and was shorn of the many nonresearch functions that had accreted in OERI over the years. Further, it was given a clear statutory mission to conduct, support, disseminate, and promote the use of scientifically valid research.

3. IES, ESEA, AND THE STATE OF EDUCATION RESEARCH

Was the National Academies wrong in 1999 about the condition of education research? If not, how could NCLB have required states and local education

agencies to use the results of scientific research for virtually every activity under the act? A partial list of NCLB provisions that require the use of scientific research includes:

- Methods and instructional strategies for comprehensive school reform;
- Technical assistance to schools;
- Professional development in schools;
- Implementation of new curriculum;
- Strengthening the core academic program of schools;
- Dissemination of information about effective school library media programs;
- Methods for student learning, teaching, and school management;
- State reform activities;
- Math and science partnerships;
- Promotion of teaching skills for mathematics and science teachers;
- Reading instruction;
- Distance learning programs for mathematics and science teachers;
- Teaching limited English proficient children;
- Prevention of illegal drug use and violence;
- Programs for effective parent and community involvement;
- Award programs for nationally significant programs;
- Character education programs;
- Meeting the educational needs of gifted and talented students;
- Dissemination of information on student achievement and school performance to parents and the community; and
- Improvement of educational opportunities for Indian children.

If we can put aside the possibility of a miraculous explosion of research findings in the brief period between the issuance of the National Academies' report in 1999 and the passage of NCLB in 2001, we have to conclude that the research base was woefully inadequate to support the demands placed on states and school districts by NCLB. It was as if the U.S. Congress had passed a law in 1931 requiring every hospital in the nation to use the results of scientific research to eliminate all cases of infectious disease, with a 1944 deadline for a zero incidence of infection. It is also a bit like growing food by decree in the old Soviet Union.

Two intertwined consequences of making people accountable for accomplishing goals they do not know how to accomplish is that they will try to get smarter, at least initially, and they will game the system. Examples of trying to get smarter in the context of NCLB include the "what works" questions I got from practitioners, the pressure on IES from the department's leadership to

identify "best practices" for reaching the NCLB goals, the creation of numerous federal technical assistance centers supposedly translating research to practice, and the audience for the dizzying array of "expert" papers and presentations that addressed how to meet the student performance requirements of NCLB. Examples of gaming the system include lowering the cut scores on statewide assessments that define student proficiency, teaching answers to specific test questions known to be likely to occur on a statewide test, focusing instructional attention on "bubble children" (i.e., students within short reach of the scores needed to be deemed proficient on the state test), and outright falsification of test scores.

The Education Sciences Reform Act, in keeping with its title and its intent, provided a definition of scientific research that was to guide the work of IES and distinguish it from what had become the dominant forms of education research in the latter half of the twentieth century—qualitative research grounded in postmodern philosophy and methodologically weak quantitative research. The historical trend in education research away from the canons of quantitative science has been amply documented.

One window into this trend is the decline in studies designed to measure the effectiveness of education programs and practices. One of my first initiatives after taking office was to commission a survey of education practitioners to determine what they wanted from education research. Their number one priority was research on what works to raise student achievement in reading, math, and science. Whereas questions of what works are paramount to educators, there was declining interest in those questions in the education research community prior to IES.

Quantitative research on program effectiveness was frequently replaced by activities in the tradition of postpositivism and deconstructivism in the humanities. These approaches are based on philosophical assumptions that question the existence of a physical reality beyond what is socially constructed—for example, "Another type of scientificity is needed for the social sciences, a postpositivist, interpretive scientificity that takes into account the ability of the object to object to what is told about it" (Lather 2007, p. 71). (Translation: what social scientists conclude about people has to be based on people's beliefs.)

Even those portions of the education research community committed to empiricism all too frequently deployed research designs that could not support causal conclusions, while drawing such conclusions with abandon. Examples of weak methods paired with strong conclusions in education research abound. For example, the Northwest Regional Educational Laboratory carried out an evaluation of the impact of Reading First, NCLB's signature national reading intervention, and concluded that the program was having a positive impact (Deussen, Nelsestuen, and Scott 2008). The study, and many others trumpeted

by advocates of Reading First and other favored education programs, involved no comparison group or credible counterfactual of any sort. The conclusion of a positive impact was based entirely on test scores rising in Reading First schools.

That determining the impact of a variable requires a credible comparison condition is so fundamental to the quantitative social and behavioral sciences that it hardly requires mention. It is shocking that so much of what passes as education research—even when numbers are being crunched—does not incorporate this fundamental requirement of causal reasoning.

4. EXPERIMENTS AS A GOLD STANDARD FOR CERTAIN QUESTIONS

In the context of declining interest in studies of the effectiveness of education programs, the ascendance of postmodern approaches to education research, and the frequent use of weak methods to support strong causal conclusions, IES took a clear stand that education researchers needed to develop interventions that were effective in raising student achievement and to validate the effectiveness of those interventions using rigorous methods (as defined and accepted within the quantitative social, behavioral, economic, cognitive, and health sciences).

The best and most rigorous methods vary with the question being addressed. They include methods for producing sound descriptive summaries, including surveys, assessments, observational data, and administrative records; methods appropriate for generating and refining hypotheses about possible causal relationships, such as multivariate analysis; methods appropriate to the development of new interventions, such as testing prototypes for usability; and methods designed to address questions concerning the effectiveness of particular policies or practices, including single-subject, quasi-experimental, and experimental approaches.

For questions about the effectiveness of particular policies and practices (that is, what works), randomized field trials provide the most reliable answers. The methodological superiority of randomized trials for drawing causal claims in areas in which outcomes are affected by many variables and in which effects vary across individuals and settings is very widely acknowledged across all the sciences, including education. In education, the National Research Council's report, *Scientific Research in Education*, concludes that "nonrandomized studies are weaker in their ability to establish causation than randomized field trials, in large part because the role of other factors in influencing the outcome of interest is more difficult to gauge in nonrandomized studies" (National Research Council 2002, p. 110). A follow-up report from a second National Research Council committee concurs that the randomized trial is the best design for making causal inferences about the effectiveness of educational programs

and practices (National Research Council 2004). Similarly, a report by the American Educational Research Association concludes that "the statistical solution to the fundamental problem of causality relies on the assumption of independence between pretreatment characteristics and treatment group assignment. This independence is difficult to achieve in nonrandomized studies. . . . This is why randomized field trials are often considered the 'gold standard' for making causal inferences" (Schneider et al. 2007, p. 16).

Another desirable characteristic of randomized trials is that they are easy to interpret. In a well-designed and implemented trial, the impact of the intervention is the difference in the post-test means of the intervention and control group expressed in a natural unit—for example, earnings in dollars, or in some form of effect size. Further, the threats to the internal validity of trials, such as differential attrition, are well understood and can be reduced to rules for judging studies that can be applied reliably by trained raters—for example, as they have been by the IES's What Works Clearinghouse (see http://ies.ed.gov/ncee/wwc/).

Other methodologies that are capable in principle of producing unbiased impact estimates (e.g., the use of instrumental variables, regression discontinuity designs, and multiple replications of quasi-experimental findings) require many more assumptions and more complicated statistical procedures, and they depend on the persuasiveness of the researchers' tests of whether they have specified their analytic model correctly and chosen the appropriate data to model. The result is that judgments of whether sophisticated nonexperimental approaches have been implemented sufficiently well to permit strong causal conclusions about program impact are best left to a small and elite cadre of experts who can still disagree on the merits of a particular study.

In the political arena in which decisions are made about education policy and programs—decisions that one hopes are informed by the best research—this opens the door to "you have your experts and I have mine." Of course, "experts" can be marshaled to challenge the findings from randomized field trials, but if the experts are indeed experts and the study has been well conducted, the criticisms are likely to go to issues of external validity—for example, "It is just one study in a limited number of settings." That is a very different battleground than one based on questions about internal validity—that is, valid causal inference: if there are real questions about whether a model has been specified correctly in a nonexperimental study or whether the researchers have cherry-picked the data that are most likely to support their hypothesis, the prudent skeptic can reject the study out of hand. However, real questions about the external validity of a well-conducted field trial go to how far the needle of policy action should move based on the study, not whether it should move at all. On this issue policy makers and advocates can and probably should disagree, and

they can do so without needing to reject the scientific credibility of the study itself. This gives randomized trials a superior position to nonexperimental approaches in policy debates even when the nonexperimental studies appear to have been well done and seem to allow unbiased estimates of effects.

Questions of what works are paramount for education practitioners and policy makers. Hence research investments by IES were designed to achieve the principal goal of developing or identifying a substantial number of programs, practices, policies, and approaches that enhance academic achievement and that can be widely deployed. In its research competitions, IES gave a competitive preference to randomized trials for research at the final stage of this goal, which involves a demonstration of effectiveness in practice at scale. And in its evaluations of federally supported education programs, IES deployed randomized designs whenever possible.

However, effective programs and practices do not spring forth fully formed in education any more than effective pharmaceuticals arise spontaneously in medicine. For that reason a substantial portion of IES funding supported upstream work in which researchers were developing new programs or identifying promising practices, using methods appropriate for those investigations. Another significant segment of the funding stream went to develop the infrastructure for statewide longitudinal data systems of student and teacher records. It was also used to support researchers applying nonexperimental methods to those data systems examining hypotheses about the impact of local, state, and national programs and practices (e.g., the effects of statewide class size reduction legislation or district-wide promotion policies).

5. THE YIELD

I have argued that the situation in 2001 was not so different from that described thirty years earlier for the President's Commission on Education Finance—little was known about what consistently and unambiguously makes a difference in student outcomes. This is hyperbole in the service of a broad and valid point, not an assertion that there was no useful education research prior to 2001. For example, economists launched a surge of student outcomes research in the late 1970s that accelerated in the 1990s when states like Texas began to have comprehensive administrative data and annual testing. Likewise, psychologists generated a body of intervention research in the 1980s and 1990s, to which I contributed, that is relevant to education. But these efforts were on the periphery of education research. Ten years later things have changed. Methodologically rigorous and educationally relevant education research has moved into the mainstream. Experimental methods are at the center of that progress.

How is strong research to trump weak research in a marketplace that is unsophisticated with regard to research quality? There has to be an entity that

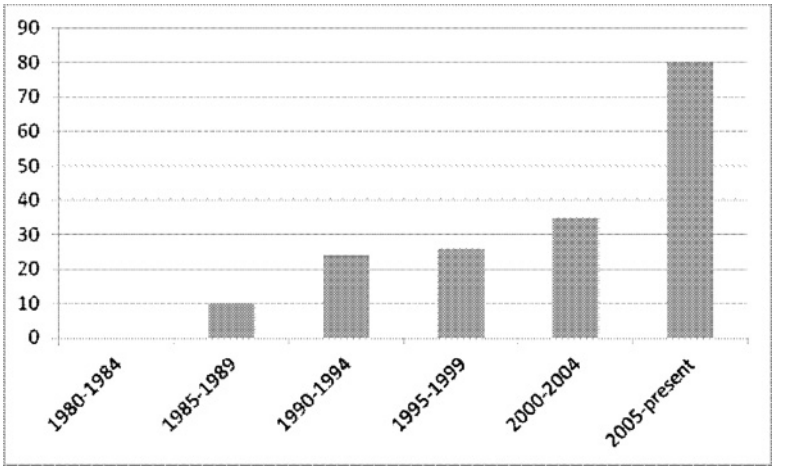

Figure 1. Number of Studies Found to Meet Evidence Standards by the What Works Clearinghouse, Grouped into Five-Year Periods by Publication Date, Beginning in 1980. Data extracted from a 3 November 2011 search on studies meeting evidence standards at http://ies.ed.gov/ncee/wwc/ResearchStudies.aspx

vets research on program effectiveness for practitioners and policy makers using rigorous scientific standards. And it has to become the preeminent source for such information, effectively muting the cacophony of conflicting claims and assertions that arise from those who advocate with numbers or draw conclusions based on methods that cannot support causal conclusions. The Food and Drug Administration serves this function in the marketplace for therapeutic pharmaceuticals and has had a transforming effect on health outcomes in the United States and the world by elevating science over quackery, opinion, and professional best (and often wrong) guess.

Enter the What Works Clearinghouse (WWC), which was first funded by IES in 2002 with the goal of being the central and trusted source of scientific evidence for what works in education. Limited in its focus to well-implemented experiments and quasi-experiments with pretest equating that examine the impact of education programs and practices, the WWC has identified through systematic literature reviews eighty-eight separate interventions with positive and potentially positive evidence of effectiveness across the domains of academic achievement, dropout prevention, language development, mathematics/science achievement, personal/social development, and reading/writing. Going from nothing known about what makes a difference to eighty-eight research-proven and available education interventions is considerable progress.

The upswing in the number of studies that address important education issues with well-designed and implemented experimental methods is illustrated in figure 1, which presents the publication dates of WWC-reviewed studies that

meet evidence standards, grouped by five-year periods beginning in 1980. The number of such studies published over the past ten years is roughly double the number published in the preceding twenty years.

Another area in which the yield from experiments is impressive is the evaluation of federal education programs. In 2000 the Department of Education had eighty-four program evaluations and studies under way, only one of which was a randomized trial (Mosteller and Boruch 2002). When IES took over the responsibility for most of the department's impact evaluations, it committed to using randomized field trials whenever possible. In fact, over twenty trials have been launched in the last decade, and the few evaluations that were unable to use randomization used the next best thing, regression discontinuity.

The yield from these studies tilts more toward identifying what does not work than what does work. For example, several studies of what seemed to be promising approaches to teacher professional development in reading and mathematics failed to identify any impact on student achievement (although they did affect teacher practice). In addition, several big-ticket line item education programs were found to have no overall impact on student outcomes, including Even Start (a preschool family literacy program), Reading First (an intensive reading intervention for early elementary school), and Upward Bound (a program to increase the college-going and completion rates among low-income and minority students in high school). But there were also notable findings of positive impacts, including increased high school graduation rates for low-income students in the District of Columbia who were given access to vouchers to attend private schools, improved mathematics achievement in second grade for students exposed to Saxon Math (as compared with three other commercially available curricula), and structured English immersion that produced larger gains on tests of academic achievement in English than a transitional program involving bilingual instruction in Spanish and English.

Congress and the executive branch have taken the results from these rigorous evaluations very seriously in decisions about program funding, although not always in the expected ways. Reading First and Even Start were both eliminated, with the evaluation results front and center in the debate. However, in reauthorizing Upward Bound, Congress rejected the evaluation findings and stipulated that the program could never again be subject to a rigorous impact evaluation.

6. REMAINING CRITICISM

While all observers agree that things have changed, not everyone agrees that all the changes have been desirable. One category of criticism rejects the role of experiments in advancing the practice of education. The argument is for the

exceptionality of education as a field subject to the canonical methods of the social and behavioral sciences.

One version of this perspective is that every child, teacher, and school is different from every other child, teacher, and school, thus preventing generalizable findings on what works. Another version is that there are so many variables affecting student learning that nothing other than systemic and wholesale change that simultaneously encompasses schools, teachers, homes, and communities will ever produce effects that are large enough to make a difference.

> In the past—indeed, in the present—much of the best school practice has been based on . . . seat-of-the-pants observations, reflections, and informal experimentation. Perhaps we need to be doing more of this, rather than less; perhaps, in fact, research dollars might be better spent on setting up teacher study groups or mini-sabbaticals, rather than on NIH-style field-initiated or targeted-grant competitions. (Gardner 2002, p. 72)

I would recommend to those who express such views the study of the history of other fields that have gone through a transformation to evidence-based practice. The arguments they present against the applicability of the scientific method to education were, for example, prevalent in medicine in the first part of the twentieth century (Marks 2000).

Another criticism is that the research methods appropriate to education are much broader than "gold standard" randomized field trials. Of course that is correct. The challenge is always to match the method to the question. The methods appropriate to developing and validating a new generation of academic assessments, for example, are much different than the methods appropriate to determining whether charter schools have an impact on student achievement. This has always been well recognized in federal funding for education research.

IES has five research goals for its research programs:

1. Identification: Identify existing programs, practices, and policies that may have an impact on student outcomes and the factors that may mediate or moderate the effects of these programs, practices, and policies;
2. Development: Develop programs, practices, and policies that are theoretically and empirically based;
3. Efficacy: Evaluate the efficacy of fully developed programs, practices, and policies;

4. Scale-up: Evaluate the impact of programs, practices, and policies implemented at scale; and
5. Measurement: Develop and/or validate data and measurement systems and tools.

IES funding announcements strongly encourage randomized trials only for applications under the efficacy and scale-up goals, where the intent is to draw causal inferences about program impact. The language in the funding announcements for the other goals very clearly indicates that other methodological approaches are desired. The identification goal prioritizes the statistical modeling of observation data. The development goal prioritizes the collection of empirical data that will provide feedback for refining successive prototypes of the intervention that is being developed. IES does not support applications under the development goal that involve testing the efficacy of interventions in a significant number of classrooms or schools using randomized experiments. The measurement goal is about assessments, and the appropriate methods are psychometric, involving demonstrations of validity and reliability.

If IES does not limit its funding to research employing randomized trials (and it does not), what is the explanation for the frequency with which IES is criticized for its "narrow methodological focus"? I believe it is by and large nothing but opposition from those whose approach to education research falls outside the accepted canons of the social and behavioral sciences. These are researchers who speak of qualitative research methods not as one among a large number of research tools that should be in the researcher's tool kit, to be deployed as appropriate to the question being asked, but as a way of thinking and a category of scholar, as in, "I am a qualitative researcher and I don't get funding from IES." These are academics who frequently refer dismissively to their colleagues in economics, psychology, or epidemiology as number crunchers. These are the individuals who advocate with numbers, whose conclusions are drawn before they begin the search for consistent data. These are the individuals who were responsible for, in the words of the National Academy of Sciences, education being like no other field in the research base being so inadequate and little used.

7. THE FUTURE OF EXPERIMENT

There has assuredly been considerable progress within education in the application of methods that support strong causal reasoning, but there is much still to accomplish. I do not mean just in the sense that science is never complete and thus the nation will always need a vibrant education research enterprise. Rather, I mean aspects of the present endeavor that are categorical deficiencies

in the sense that they are manifestly present in research in other fields but generally lacking in education research. These include the following.

Research That Advances Conceptual Models

The present yield from rigorous experiments in education as represented in the WWC is largely a catalog of effect sizes associated with particular branded programs and practices. Knowing that intervention A has positive evidence of effectiveness with respect to academic outcomes, whereas intervention B does not, divulges little or nothing about the key processes underlying learning or the key ingredients of a successful program. Contrast this with the deep theoretical knowledge and bench research that drives the development of pharmaceuticals. Pharmaceutical A never gets to a field trial without that basis. Results from a drug trial that are smaller than anticipated or that are accompanied by unanticipated side effects often lead to a redesign that is intelligent, in the sense that it is driven by the underlying biochemical science. Education interventions and the trials that test them are very different. All that is learned from a failure is that the hope was misplaced. The closest that education currently comes to conceptually driven, design-relevant research is in the cognitive science of learning and instruction. We need much more such research for the field to mature and for the nation to obtain more value from its investments in large-scale trials.

Strong Links between Trials and Administrative Data

Schools, school districts, and states will have the opportunity to be data-driven education managers as a new generation of longitudinal data systems comes on line. For the first time, states should be able to link individual students' longitudinal progress on assessments of achievement to a variety of system variables over which they have control.

Consider curriculum. Research has demonstrated that curricula can have a significant impact on student learning, suggesting that using more effective curricula could improve student outcomes at a fraction of the cost of other interventions, such as class size reductions or teacher bonuses. But the relative effectiveness of most curricula in use is unknown, so educators have no strong evidence base to use when choosing curricula. Most curricula in use have never been subjected to an impact evaluation, much less a rigorous randomized trial. The WWC has to date examined seventy-three elementary school mathematics curricula. Of these, sixty-six either have no studies of their efficacy or have no studies that meet reasonable evidence standards.

Even when a curriculum has been subjected to a well-designed trial and is found to have positive effects, the implications for practice may be unclear.

With very few exceptions, these studies will have involved business-as-usual controls and thus do not provide adequate support for typical decision making by education administrators who have to choose among multiple curricula. The WWC may reveal that curriculum A, studied in the Milwaukee schools, increases math achievement in Milwaukee compared with the instructional approaches otherwise used in Milwaukee. This information has value, but the director of curriculum and instruction in the Wake County, North Carolina, public schools cannot determine on that basis whether typical practice in Milwaukee prior to the introduction of the new curriculum was similar to typical practice in Wake County. It may be that what Wake County is already doing is better than either curriculum A as used in Milwaukee or business as usual in Milwaukee.

What it would take to make the results more meaningful for practice is a bridge between the effects that have been found in well-designed trials and the academic growth that an education practitioner or policy maker is able to see for students using administrative data. This requires developing a set of normative academic achievement growth rates for students of various characteristics and linking those rates to effective curricula demonstrated in field trials. Thus the administrator in Wake County would be able to (1) enter the demographic characteristics of third graders in that school district and have returned the expected growth in math achievement for similar students in the nation or state; (2) search a what-works database for experimentally validated math curricula proven to produce larger growth rates than those expected normatively for students similar to those in Wake County; and (3) assess the success of the implementation of the new math curriculum chosen for Wake County with respect to the historical trends in Wake County prior to adoption of the new curriculum, present national or state norms, and expectations based on the results from the what-works database. The researcher or developer could use the administrative data in other ways to support experimental trials, for example, by screening the data statewide to look for associations between particular curricula and faster than average rates of academic growth, by choosing sites for trials based on sampling that is informed by normative data on progress, or by conducting exploratory quasi-experiments that take advantage of discontinuities in the sequence of adoption of particular curriculum materials.

Funding

Although there have been respectable increases in federal funding for education research in relative terms, the absolute levels of present funding are paltry, representing less than 1 percent of the total discretionary budget of the

U.S. Department of Education. In contrast, the U.S. Department of Health and Human Services devotes more than 40 percent of its discretionary budget to knowledge creation through the National Institutes of Health, the Centers for Disease Control and Prevention, and many other agencies. In research the nation gets what it pays for. Funding research for knowledge creation is a uniquely federal role. It is time to spend more on education research.

Knowledgeable Consumers

The colleges and universities that produced a generation of education researchers who think in postmodern terms and dismiss number crunching and randomized trials as historical artifacts also produce the teachers and administrators who are, ideally, the consumers of education research. But those teachers and administrators have almost no grounding in the methods of the behavioral, cognitive, social, and economic sciences. Their decision making on education policy and practice is inordinately driven by intuition and interpersonal influence. And there is little in their work environment that corresponds to tort actions and insurance reimbursement policies in the medical arena—that is, few consequences for ignoring strong scientific findings in practice. Imagine how much more we would know in education about what works, for whom, and under what circumstances if most state- and district-level education decision makers understood the value of systematically staggering the introduction of a new practice while collecting administrative data on its effects.

Institutions that train education practitioners and states that license and certify them should have to demonstrate that those practitioners have at least the requisite book knowledge to use and interpret data. Such training is largely absent.

One of the major story lines of the growth of civilization is the advance of the experiment. From the food we eat to the diseases we conquer to our understanding of how we think and behave, we have depended on an approach that marries our models of the world with tests of their validity through systematic variation to determine cause and effect. These same tools of thought are no less critical to the advance of education than to medicine, agriculture, psychology, or transportation.

REFERENCES

Averch, Harvey A., Stephen J. Carroll, Theodore S. Donaldson, Herbert J. Kiesling, and John Pincus. 1972. *How effective is schooling? A critical review and synthesis of research findings*. RAND Corporation. Available www.rand.org/pubs/reports/2006/R956.pdf. Accessed 21 November 2011.

Deussen, Theresa, Kari Nelsestuen, and Caitlin Scott. 2008. *Does Reading First work? Data trends from evaluations in five western states*. Portland, OR: Northwest Regional Educational Laboratory.

Gardner, Howard. 2002. The quality and qualities of educational research. *Education Week* 22(1): 49–72.

Lather, Patti. 2007. *Getting lost: Feminist efforts toward a double(d) science*. Albany, NY: SUNY Press.

March, James G. 1978. Foreword. In Lee Sproull, Stephen Weiner, and David Wolf. *Organizing an anarchy: Belief, bureaucracy, and politics in the National Institute of Education*. Chicago: University of Chicago Press.

Marks, Harry M. 2000. *The progress of experiment: Science and therapeutic reform in the United States, 1900–1990*. Cambridge, UK: Cambridge University Press.

Mosteller, Frederick, and Robert Boruch. 2002. *Evidence matters: Randomized trials in education research*. Washington, DC: Brookings Institution Press.

National Research Council. 1999. *Improving student learning: A strategies plan for education research and its utilization*. Washington, DC: National Academies Press.

National Research Council. 2002. *Scientific research in education*. Washington, DC: National Academies Press.

National Research Council. 2004. *Advancing research in education*. Washington, DC: National Academies Press.

President's Commission on School Finance. 1971. *Progress report of the President's Commission on School Finance*. ERIC document ED058643. Washington, DC: U.S. Department of Education.

Rudalevige, Andrew. 2008. Structure and science in federal education research. In *When research matters: How scholarship influences education policy*, edited by Frederick Hess, pp. 17–40. Cambridge, MA: Harvard Education Press.

Schneider, Barbara, Martin Carnoy, Jeremy Kilpatrick, William H. Schmidt, and Richard J. Shavelson. 2007. *Estimating causal effects using experimental and observational designs*. Washington, DC: American Educational Research Association.

Vinovskis, Maris A. 2001. *Revitalizing federal education research and development*. Ann Arbor, MI: University of Michigan Press.