Mid-Western Educational Researcher

Volume 18 | Issue 1 Article 3

2005

Evidence-Based Reform in Education: Promise and Pitfalls

Robert E. Slavin
Johns Hopkins University

Keynote Address

Evidence-Based Reform in Education: Promise and Pitfalls

Robert E. Slavin
Johns Hopkins University

Abstract

In this keynote address presented at the Mid-western Educational Research Association Annual Meeting in October, 2004, the author discusses the increasing interest of federal policy-makers in scientifically-based research. A comparison between education and other disciplines is offered, and a proposal for increased rigor in educational research is proposed.

Education is on the brink of a scientific revolution that has the potential to profoundly transform policy, practice, and research. Consider the following:

- In 1998, Congress appropriated $150 million per year to provide schools funds to adopt "proven, comprehensive reform models." This unprecedented legislation, introduced by Congressmen David Obey and John Porter, defined "proven" in terms of experimental-control comparisons on standards-based measures. To my knowledge, this was the first time in history that education funding anywhere has been linked directly to evidence of effectiveness (see Slavin, 1997). Comprehensive School Reform (CSR) funding progressively increased to $310 million annually, and has provided funding to more than 3000 mostly high-poverty schools.
- The Bush administration's main education initiative, No Child Left Behind, took the idea of scientifically-based practice to an even higher level. The No Child Left Behind legislation refers to "scientifically-based research" 110 times. It defines "scientifically-based research" as "rigorous, systematic and objective procedures to obtain valid knowledge," which includes research that "is evaluated using experimental or quasi-experimental designs…," preferably with random assignment. "Scientifically-based research" is intended to serve as the basis for a wide array of federally funded programs, especially Reading First programs for reading in grades K-3.
- Grover Whitehurst, the current director of the Institute of Education Science (IES) in the U.S. Department of Education, has taken a strong line in support of randomized experiments (Whitehurst, 2002). The U.S. Department of Education strategic plan for 2002-2007 anticipates having 75% of all OERI-funded research that addresses causal questions use random assignment designs by 2004 (previously, such research was less than 5% of causal research funded by The U.S. Department of Education). As a direct result, Congress significantly increased funding for education research. Research involving random assignment is now under way on early childhood programs, elementary and secondary reading, math, programs for English language learners, teacher professional development, after school remedial programs, and much more.

It is important to note that none of these policy developments have yet produced the revolution I am anticipating. These initiatives are too new to have had any impact on practice. Yet these and other developments, if not yet proven, still create the potential for changes with far-reaching consequences. It is possible that these policy reforms could set in motion a process of research and development on programs and practices affecting children everywhere. This process could create the kind of progressive, systematic improvement over time that has characterized successful parts of our economy and society throughout the 20th century, in fields such as medicine, agriculture, transportation, and technology. In each of these fields, processes of development, rigorous evaluation, and dissemination have produced a pace of innovation and improvement that is unprecedented in history (see Shavelson & Towne, 2002). These innovations have transformed the world. Yet education has failed to embrace this dynamic, and as a result, education moves from fad to fad. Educational practice does change over time, but the change process more resembles the pendulum swings of taste characteristic of art or fashion (think hemlines) rather than the progressive improvements characteristic of science and technology (see Slavin, 1989).

Welcome to the 20th Century

At the dawn of the 21st century, education is finally being dragged, kicking and screaming, into the 20th century.

Portions of this paper are adapted from Slavin, R.E. (2003), Evidence-based policies: Transforming educational practice and research. *Educational Researcher, 31* (7), 15-21. This paper was written under funding from the U.S. Department of Education (Grant No. OERI-R-117-D40005). However, any opinions expressed are those of the author and do not necessarily represent Department of Education positions or policies.

The scientific revolution that utterly transformed medicine, agriculture, transportation, technology, and other fields early in the 20th century almost completely bypassed the field of education. If Rip Van Winkle had been a physician, a farmer, or an engineer in the 19th century, gone to sleep, and awoke today, he would be unemployable. If he had been a good primary school teacher in the nineteenth century, he'd probably be a good primary school teacher today. It's not that we haven't learned anything since Rip Van Winkle's time. It's that applications of the findings of educational research remain haphazard, and that evidence is respected only occasionally, and only if it happens to correspond to current educational or political fashions.

Early in the 20th century, the practice of medicine was at a similar point. For example, research had long since identified the importance of bacteria in disease, and by 1865 Joseph Lister had demonstrated the effectiveness of antiseptic procedures in surgery. In the 1890s, William Halsted at Johns Hopkins University introduced rubber gloves, gauze masks, and steam sterilization of surgical instruments, and demonstrated the effectiveness of these procedures. Yet it took thirty years to convince tradition-bound physicians to use sterile procedures. If he dropped his scalpel, a physician in 1910 was as likely as not to give it a quick wipe and carry on.

Today, of course, the linkage between research and practice in medicine is so tight that no physician would dream of ignoring the findings of rigorous research. Because medical practice is so closely based on medical research, funding for medical research is vast, and advances in medicine take place at breathtaking speed. My father's cardiologist recommended that he wait a few years to have a necessary heart valve operation because he was sure that within that short span of time, research would advance far enough to make the wait worthwhile. As it turned out, he was right.

The most important reason for the extraordinary advances in medicine, agriculture, and other fields is the acceptance by practitioners of evidence as the basis for practice. In particular, it is the randomized clinical trial, more than any single medical breakthrough, that has transformed medicine (Doll, 1998). In a randomized clinical trial, patients are assigned at random to receive one treatment or another, such as a drug or a placebo. Because of random assignment, it can be assumed with an adequate number of subjects that any differences seen in outcomes are due to the treatment, not to any extraneous factors. Replicated experiments of this kind can establish beyond any reasonable doubt the effectiveness (or lack thereof) of treatments intended for applied use (see Boruch, 1997).

Experiments in Education

In education, experiments are not uncommon, but they are usually brief, artificial experiments on topics of theoretical more than practical interest, often involving hapless college sophomores. Far more rare are experiments evaluating treatments of practical interest studied over a full school year or more. I write an educational psychology textbook (Slavin, 2003) that is full of research findings of all kinds, findings that are valuable in advancing theory and potentially valuable to teachers in understanding their craft. Yet the brief experiments, correlational studies, and descriptive studies that yield most of the information presented in my text or any other educational psychology text do not collectively add up to school reform. They are suggestions about how to think about daily teaching problems, not guides to the larger questions educators and policymakers must answer. Imagine that research in cardiology described heart function and carried out small scale laboratory studies, but never developed and tested an artificial heart valve. If this were the case, I'd be an orphan. Imagine that agricultural research studied plant growth and diseases, but never developed and tested new disease-resistant crops. Educational research has produced many rigorous and meaningful studies of basic principles of practice, but very few rigorous studies of programs and practices that could serve as a solid base for policy and practice, and has had little respect for the studies of this kind that do exist. Because of this, policy makers 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 policy makers would feel to be valuable, making these policy makers unwilling to invest in large-scale, definitive research.

Shifting Policy Perspectives

The dramatic changes in federal education policies I mentioned earlier could potentially reverse this declining spiral. If the new funding flowing into research in the U.S. can produce some notable successes, we could have an ascending spiral: rigorous research demonstrating positive effects of replicable programs on important student outcomes would lead to increasing funding for such research which would lead to more and better research and therefore more funding. More importantly, millions of children would benefit in the fairly near term. Once we establish replicable paradigms for development, rigorous evaluation, replication, and dissemination, these mechanisms could be applied to any educational intervention or policy. Imagine that there were programs under way all the time to develop, evaluate, and disseminate new programs in every subject and every grade level, as well as programs on school-to-work transitions, special education, gifted programs, dropout prevention, programs for English language learners, race relations programs, drug abuse prevention, violence prevention, and so on. Every one of these areas lends itself to a development-evaluation-dissemination paradigm, as would many more. Over time, each area would experience the step-by-step, irreversible progress characteristic of medicine and agriculture, because innovations would be held to strict standards of evaluation before being recommended for wide scale use.

Research Designs

The scientific revolution in education will only take hold and produce its desired impacts if research in fact begins to focus on replicable programs and practices central to education policy and teaching, and if it in fact employs research methods that meet the highest standards of rigor.

This begs an important question: What kinds of research are necessary to produce findings of sufficient rigor to justify faith in the meaning of their outcomes?

Of course, all sorts of research designs are appropriate for various purposes, from description to theory building to hypothesis testing. However, leaders in the current administration and many other educational researchers throughout the world (see Angrist, 2004) have been arguing that nothing less than *randomized* experiments will do for evaluations of educational interventions and policies. When we want to know the outcome of choosing program X instead of program Y, there is no equivalent substitute for a randomized experiment.

Randomized experiments

The difference in the value of randomized and well-matched experiments relates primarily to the problem of selection bias. In a matched experiment, it is always possible that observed differences are due not to treatments, but to the fact that one set of schools or teachers was willing to implement a given treatment while another was not, or that a given set of students selected themselves or were selected into a given treatment while others were not.

When selection bias is a possibility at the student level, there are few if any alternatives to random assignment, because unmeasured (often, unmeasurable) pre-existing differences are highly likely to be alternative explanations for study findings. For example, consider studies of after school or summer school programs. If a researcher simply compared students attending such programs to those not attending who were similar in pretest scores or demographic factors, it is very likely that unmeasured factors such as student motivation, parents' support for education, or other consequential factors could explain any gains observed, because the more motivated children are more likely to show up. Similarly, studies comparing children assigned to gifted or special education programs to students with similar pretest scores are likely to miss key selection factors that were known to whoever assigned the students but not measured. If one child with an IQ of 130 is assigned to a gifted program and another with the same IQ is not, it is likely that the children differ in motivation, conscientiousness, or other factors. In these kinds of situations, use of random assignment from within a selected pool is essential.

In contrast, there are situations in which it is teachers or schools that elect to implement a given treatment, but there is no selection bias that relates to the children. For example, a researcher might want to compare the achievement gains of children in classes using cooperative learning, or schools using comprehensive reform models, to the gains made by control groups. In such cases, random assignment of willing teachers or schools is still far preferable to matching, as matching leaves open the possibility that volunteer teachers or staffs are better than non-volunteers. However, the likely bias is much less than in the case of student self-selection. Aggregate pretest scores in an entire school, for example, should indicate how effective the current staff has been up to the present, so controlling for pretests in matched studies of existing schools or classes would control out much of the potential impact of having more willing teachers. For external validity, it is crucial to note that the findings of a well-matched experiment comparing volunteers to non-volunteers apply only to schools or teachers who volunteer, but the potential for bias is moderate (after controlling for pretests and demographic factors).

The importance of this discussion lies in the fact that randomized experiments of interventions applying to entire classrooms can be extremely difficult and expensive to do, and are sometimes impossible. My colleagues and I at Johns Hopkins University are doing a randomized evaluation of Success for All, a comprehensive reform model. Recruiting schools for this study was extremely difficult, even though we are offering substantial financial incentives to schools willing to be assigned at random to experimental or control groups. For the cost of doing this randomized study, we (and others) could have done two or three equally large-scale matched studies. It is at least arguable that replicated matched studies, done by different investigators in different places, might produce more valid and meaningful results than one definitive, once-in-a-lifetime randomized study.

Still, fully recognizing the difficulties of randomized experiments, I think they are nevertheless possible in most areas of policy-relevant program evaluation, and whenever they are possible, they should be used. Reviews of research in other fields have found that matched studies generally find stronger outcomes than randomized studies, although usually in the same direction (e.g., Friedlander & Robins, 1995; Fraker & Maynard, 1987; Ioannidis et al, 2001). Four randomized experiments we are doing at Johns Hopkins University and the Success for All Foundation illustrate the potential and the pitfalls. One of these, which I mentioned earlier, involves randomly assigning 41 schools to Success for All or control conditions for a 3-year experiment. Initially, we offered $30,000 to each school, but we got hardly any takers. Schools were unwilling to take a chance on being assigned to the control group for three years.

In spring, 2002, we changed our offer. Schools willing to participate were randomly assigned to use Success for All either in grades K-2 or in 3-5. Recruitment was still difficult, but under this arrangement, we signed up adequate numbers of schools.

For another study led by my colleague Bette Chambers, we recruited schools for a third-party study of the Curiosity

Corner preschool model. We offered schools the program for free, to start either in 2003-2004 or 2004-2005 (with random assignment to the two start dates). The 2004-2005 group serves as the control group in 2003-04. This delayed treatment control group design was easy for schools to accept, and we did not have serious recruiting problems. We're doing a nearly identical study of an after-school program, and again, recruitment was not difficult.

We recently completed a study of the use of embedded multimedia, video vignettes embedded in beginning reading instruction (Chambers et al., 2004). Again, ten schools were randomly assigned to receive the multimedia materials immediately or one year later. Finally, my colleague Geoff Borman did randomized evaluations of summer school programs, in which individual children were randomly assigned to participate now or later (Borman, Boulay, Kaplan, Rachuba, & Hewes, 2001). In all of these cases, obtaining sufficient volunteers was not difficult.

These examples of a diverse set of research problems illustrate that one way or another, it is usually possible to use random assignment to evaluate educational programs. There is no one formula for randomization, but with enough resources and cooperation from policy makers, random assignment is possible.

Beyond the benefits for reducing selection bias, there is an important *political* reason to prefer randomized over matched studies at this point in history. Because of political developments in the U.S., we have a once in a lifetime opportunity to reverse the "awful reputation" that educational research has among policy makers (Kaestle, 1993; Lagemann, 2002). This is a time when it makes sense to concentrate resources and energies on a set of randomized experiments of impeccable quality and clear policy importance, to demonstrate that such studies can be done. Over the longer run, I believe that a mix of randomized and rigorous matched experiments evaluating educational interventions may be healthier than a steady diet of randomized experiments, but right now we need to establish the highest possible standard of evidence, on a par with standards in other fields, to demonstrate what educational research can accomplish.

Non-Experimental Research

I should hasten to say again that forms of research other than experiments, whether randomized or matched, can also be of great value. Correlational and descriptive research are essential in theory building and in suggesting variables worthy of inclusion in experiments. Our Success for All program, for example, owes a great deal to correlational and descriptive process-product studies of the 1970's and 1980's (see Slavin & Madden, 2001). As components of experiments, correlational and descriptive studies can also be essential in exploring variables that go beyond overall program impacts. In some policy contexts, experiments are impossible, and well-designed correlational or descriptive studies may be sufficient.

The experiment, however, is the design of choice for studies that seek to make causal conclusions, and particularly for evaluations of educational innovations.

Basing Educational Policy on Evidence

Historically, the impact of education research on education practice has been tenuous at best. Innovation takes place, but it is based on fads and politics rather than evidence. At best, education policies are said to be "based on" scientific evidence, but are rarely scientifically evaluated. This distinction is critical. The fact that a program is based on scientific research does not mean that it is in fact effective. For example, imagine an instructional program whose materials are thoroughly based on scientific research, but which is so difficult to implement that in practice, teachers do a poor job of it, or which is so boring that students don't pay attention, or which provides so little or such poor professional development that teachers do not change their instructional practices. Before the Wright brothers, many inventors launched airplanes that were based on exactly the same "scientifically-based aviation research" as the Wright brothers used at Kitty Hawk, but the other airplanes never got off the ground. Worse, any program or policy can find some research somewhere that suggests it might work.

Given the current state of research on replicable programs in education, it would be difficult to require that government funds be limited to programs that have been rigorously evaluated, because there are so few such programs. However, programs that do have strong, rigorous evidence of effectiveness should be emphasized over those that are only based on valid principles, and there needs to be a strong effort to invest in development and evaluation of replicable programs in every area, so that eventually legislation can focus not on programs "*based* on scientifically-based research" but on programs that have actually been successfully evaluated in rigorous experiments.

Research Syntheses

The evidence-based policy movement is by no means certain to succeed. Education has a long tradition of ignoring or even attacking rigorous research. Researchers themselves, even those who fundamentally agree on methodologies and basic principles, may disagree publicly about the findings of research. These disagreements, which are a healthy and necessary part of the scientific process, will be seized upon by individuals who oppose the entire concept of evidence-based reform as indications that even the experts disagree.

For these and many other reasons, it is essential that independent review commissions representing diverse viewpoints be frequently constituted to review the research and produce consensus on what works, in language that all educators can access. In the area of reading, it is impossible to overstate the policy impact of the National Research Council (Snow, Burns, & Griffin, 1998) and National Reading Panel (1999) reports, which produced remarkable consensus on the state of the evidence in early literacy. Consensus panels of this kind, with deep and talented staff support, should be in operation continually, on a broad range of policy-relevant questions, so that practitioners and policy makers can have a way to cut through all the competing claims and isolated research findings to get to the big picture findings that methodologically sophisticated researchers can agree to represent the evidence fairly and completely. The federally-funded What Works Clearinghouse is carrying out rigorous reviews of research on a range of programs and practices. This effort is just getting under way, but it could become very influential if it gives government funders a basis for favoring well-evaluated practices.

Potential Impact of Evidence-Based Policies on Educational Research

Up to now, I've spoken primarily about the potential impact of evidence-based policies on education policies and practice. I'd now like to consider the potential impact on educational research.

I believe that if evidence-based policies take hold, this will be enormously beneficial for all of educational research, not just research involving randomized or matched experiments. First, I am confident that when policymakers perceive that educational R&D is actually producing programs that are shown in rigorous experiments to improve student outcomes, they will fund research at far higher levels. This should not be a zero-sum game, in which new funds for experiments will be taken from the very limited funds now available for educational research (see Shavelson & Towne, 2002). Rather, I believe that making research relevant and important to policymakers will make them more, not less, willing to invest in all forms of disciplined inquiry in education, be it correlational, descriptive, ethnographic, or otherwise. The popularity of medical research depends totally on its ability to cure or prevent diseases, but because randomized experiments routinely identify effective treatments (and protect us from ineffective treatments), there is vast funding for *basic* research in medicine, including epidemiological, correlational, and descriptive studies. Researchers and developers will be able to argue convincingly that basic research is essential to tell us what kinds of educational programs are worth evaluating.

A climate favorable to evidence-based reform will be one in which individual researchers working on basic problems of teaching and learning will be encouraged and funded to take their findings from the laboratory or the small-scale experiment, or from the observation or interview protocol, to themselves develop and then rigorously evaluate educational treatments. Education is an applied field. Research in education should ultimately have something to do with improving outcomes for children.

Conclusion

Evidence-based policies have great potential to transform the practice of education, as well as research in education. Evidence-based policies could finally set education on the path toward progressive improvement that most successful parts of our economy and society embarked upon a century ago. With a robust R&D enterprise and government policies demanding solid evidence of effectiveness behind programs and practices in our schools, we could see genuine, generational progress instead of the usual pendulum swings of opinion and fashion.

This is an exciting time for educational research and reform. We have an unprecedented opportunity to make research matter, and to then establish once and for all the importance of consistent and liberal support for high-quality research. Whatever their methodological or political orientations, educational researchers should support the movement toward evidence-based policies, and then set to work to generate the evidence that will be needed to create the schools our children deserve.

References

Angrist, J. D. (2004). American education research changes tack. *Oxford Review of Economic Policy, 20*(2), 198-212.

Borman, G., Boulay, M., Kaplan, J., Rachuba, L., & Hewes, G. (2001). *Randomized evaluation of a multi-year summer program: Teach Baltimore.* Baltimore: Johns Hopkins University, Center for Research on the Education of Students Placed at Risk.

Boruch, R. F. (1997). *Randomized experiments for planning and evaluation: A practical guide*. Thousand Oaks, CA: Corwin.

Chambers, B., Cheung, A., Gifford, R., Madden, N., & Slavin, R. E. (2004). *Achievement effects of embedded multimedia in a Success for All reading program.* Baltimore: Success for All Foundation.

Doll, R. (1998). Controlled trials: The 1948 watershed. *British Medical Journal, 317*, 1217-1220.

Fraker, T., & Maynard, R. (1987). The adequacy of comparison group designs for evaluations of employment-related programs. *Journal of Human Resources, 22*(2), 194-227.

Friedlander, D., & Robins, P. K. (1995). Evaluating program evaluations: New evidence on commonly used nonexperimental methods. *American Economic Review, 85*(4), 923-937.

Gallagher, J. (2002). What next for OERI? *Education Week, 21*(28), 52.

Herman, R. (1999). *An educator's guide to schoolwide reform*. Arlington, VA: Educational Research Service.

Ioannidis, J. P. A. et al. (2001). Comparison of evidence of treatment effects in randomized and nonrandomized studies. *Journal of the American Medical Association, 286*(7), 821-830.

Kaestle, C. F. (1993). The awful reputation of educational research. *Educational Researcher, 22*(1), 23-26-31.

Lagemann, E. C. (2002, April). *An elusive science: The troubling history of education research*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Levin, M. E., & Levin, J. R. (1990). Scientific mnemonics: Methods for maximizing more than memory. *American Educational Research Journal, 27*, 301-321.

Linn, R. L., & Haug, C. (2002). Stability of school-building accountability scores and gains. *Educational Evaluation and Policy Analysis, 24*(1), 29-36.

Mosteller, F., & Boruch, R. (Eds.) (2002). *Evidence matters: Randomized trials in education research*. Washington, DC: Brookings.

National Reading Panel (1999). *Teaching children to read*. Washington, DC: U.S. Department of Education.

Rhine, W. R. (1981). *Making schools more effective: New directions from Follow Through.* New York: Academic Press.

Shavelson, R. J., & Towne, L. (Eds.) (2002). *Scientific research in education.* Washington, DC: National Academy Press.

Slavin, R. E. (1989). PET and the pendulum: Faddism in education and how to stop it. *Phi Delta Kappan, 70*, 752-758.

Slavin, R. E. (1997). Design competitions: A proposal for a new Federal role in educational research and development. *Educational Researcher, 26*(1), 22-28.

Slavin, R. E. (2003). *Educational psychology: Theory into practice*. (7th ed.). Boston: Allyn & Bacon.

Slavin, R. E., & Madden, N.A. (Eds.) (2001). *One million children: Success for All*. Thousand Oaks, CA: Corwin.

Snow, C. E., Burns, S. M., & Griffin, P. (Eds.). (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academy Press.

Southwest Educational Research Laboratory (2002). CSRD database of schools (web site: http://www.sedl.org/csrd/awards.html) Austin, TX: Author

U. S. Congress (2001). *No Child Left Behind Act of 2001*. Washington, DC: Author.

U. S. Department of Education (2002). *2002 application for new grants: Preschool curriculum evaluation research grant program* (CFDA No. 84.305J). Washington, DC: Author.

U. S. Department of Education (2002b). *Strategic plan, 2002-2007*. Washington, DC: Author.

U. S. Department of Education (2002c). *Draft guidance on the Comprehensive School Reform Program* (June 14, 2002 update). Washington, DC: Author.

Whitehurst, G. (2002). *Charting a new course for the U. S. Office of Educational Research and Improvement*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.