Response to Comments

Evidence-Based Reform in Education: Which Evidence Counts?

Robert E. Slavin

A key purpose I had in writing "What Works: Issues in Synthesizing Educational Program Evaluations" (this issue of *Educational Researcher,* pp. 5–14) was to encourage open discussion about how to review program evaluation research to provide unbiased, scientifically valid, and useful information that educators and policy makers can use to make wise decisions to benefit children. I was therefore delighted that the editors of *Educational Researcher* invited a diverse and distinguished group of scholars to provide commentary: Derek C. Briggs, Madhabi Chatterji, Mark Dynarski, Judith L. Green and Audra Skukauskaitė, and Finbarr Sloane. Their comments, which also appear in this journal issue, reflect very different perspectives but overlap in many ways. Rather than addressing them one at a time, I will try to draw and then discuss common themes across the five commentaries.

The Purpose of Program Evaluation Syntheses

To begin, it is important to be clear about the purpose of program evaluation syntheses, such as those produced by the What Works Clearinghouse (WWC) and the Best Evidence Encyclopedia (BEE). These syntheses are intended primarily to communicate evidence to educators making real-life choices among programs or practices. A study or set of studies is evaluated on the basis of its contribution to the knowledge that an enlightened educator would need to make sound, evidence-based choices. The syntheses do not rate the overall quality of studies and do not use the kinds of criteria that might be appropriate for publication or funding. Instead, they value external validity and practical utility as highly as research design and pay little attention to a study's contribution to theory. I recognize that there is much more worth knowing from research than what is synthesized in systematic reviews and that qualitative and descriptive research may be the best way to inform issues beyond program effects measured on tests. However, when educators ask, for example, which of two approaches is most likely to increase their students' achievement, this is a question worth answering with the best comparative, experimental evidence we can muster. Program evaluation syntheses provide useful summaries of such evidence.

Why and How Evidence-Based Reform Matters

Evidence-based reform refers to policies and practices that are explicitly based on evidence of "what works" in education. Stated in the broadest terms, it is difficult to disagree that evidence should play a role in educational practice, but there are legitimate questions about what evidence matters for practice. As noted by several of the commentators on my article, the issue is one of generalization. Many scholars, articulately represented by Green and Skukauskaitė and by Chatterji, express concern that research in education is too context bound to allow for generalizations about the likely effectiveness of educational programs. Certainly, program effects vary a great deal from study to study, and from school to school and from teacher to teacher within studies. Programs affect different students and different categories of students differently. These differences may relate to quality of implementation, characteristics of the setting, the nature of the children involved, and so on. Sometimes it is possible to find consistent interactions between treatments and school, teacher, or student characteristics on student achievement outcomes across studies, but this is often not feasible. The reviewer is left to try to find an average overall effect and then to characterize the strength of the evidence behind that average, with humility and hopes for more and better research in the future. Given that context does matter and that treatment outcomes vary, is the entire exercise worthwhile?

To answer this question, consider a thought experiment. Imagine that the principal of an elementary school with persistently low achievement levels in math asks you for advice. She must decide whether to acquire a new textbook, Textbook A, or have her teachers participate in a professional development program, Program X. You happen to know that five matched and two randomized experimental studies across a variety of school contexts have evaluated Textbook A in comparison with standard textbooks and that the average effect size was near zero. None of the studies found much difference between the achievement of students who used Textbook A and those who used other texts, as measured on standardized tests. In contrast, you know that across 7 matched and 3 randomized studies, Program X has generally had positive effects on standardized achievement tests. Outcomes vary from study to study, but overall effects are positive, and the larger, better designed studies produce the best effects.

Given this set of conditions, how should you answer the inquiring principal? If you take the view that all knowledge in education is contextual, you would have to say that you have nothing to tell her. All situations are unique, so the findings of the various studies may or may not apply to her school.

Educational Researcher, Vol. 37, No. 1, pp. 47–50
DOI: 10.3102/0013189X08315082

I would argue that such a response is irresponsible. With appropriate cautions (such as noting the need to focus on thoughtful, high-quality implementation), the research bases for the two programs are essential information that the principal should take into account in making her decision.

Now, change the thought experiment to replace the principal with a superintendent, state superintendent, or federal policy maker. The larger the number of schools to which a given policy is intended to apply, the more likely it is that any contextual differences from classroom to classroom or school to school will even out. Therefore, the average effect of many high-quality studies across diverse settings is likely to predict the average effect across many schools.

If you accept the idea that it is worthwhile to tell principals or superintendents that evidence evaluating a program they are considering supports or does not support its use, then you have accepted the core assumption behind the WWC, the BEE, and other program evaluation reviews. Effect sizes averaged across studies of various programs do not tell us everything we would want to know about the programs or about school reform in general; descriptive, qualitative, and other studies also provide valuable perspectives to educators as well as scholars. Yet well-supported averages matter. Because they matter, it is important to carefully consider how to synthesize the findings of program evaluation studies in a way that provides the best possible summary to educators and policy makers.

What Is at Stake in Evidence-Based Reform?

The importance of providing educators and policy makers with well-justified average effects of programs goes far beyond the benefit to the particular children in schools today. Such reviews play a crucial role in a broader movement toward evidence-based reform that must eventually take place in education, as it has in other fields. When education becomes an evidence-based profession, educators will routinely consult research to make choices among programs and practices. Policy makers will provide encouragement and incentives for schools to choose and effectively implement proven programs. Because of this, developers and publishers, as well as governments, will invest in innovative research and development to create new programs, technologies, and strategies capable of improving student outcomes in every subject and grade level. Knowing that programs capable of succeeding in rigorous evaluations will be favored, innovators of all kinds will come forward with ever more effective solutions to longstanding problems of education. The teaching profession will gain in stature and respect, as does every field in which professionals possess proven techniques to solve critical problems.

There are three essential requirements for evidence-based reform. First, there is a need for development and rigorous evaluation of promising innovations capable of being used on a broad scale. Second, there is a need for federal, state, and local policies to support the use of proven programs and the research and development process that produces them. Third, and perhaps most important, there is a need for systematic reviews of research that make the findings of research readily available to educators and policy makers. Making these reviews fair and well justified is the focus of my earlier article in this issue. I will now turn to the comments about how to do this.

Comments on Methodological Issues

One of the most important messages that I intended to communicate in my article was that, in light of the current state of the program evaluation literature, all reviewers are faced with an uncomfortable choice between setting very high standards and therefore including very few studies or compromising on standards and having much more to say (with appropriate caveats). However, the issues on which various reviewers choose to compromise and those on which they stand firm vary enormously, and this is what produces differences in conclusions.

Minimizing Bias

In my article, I made the argument that the guiding principle for deciding when to compromise and when to stand firm should be based on a principle of minimizing bias. Dynarski points out that the word *bias* has a specific meaning in statistics. But I meant bias in the common English sense: factors that skew conclusions in one direction or another. Dynarski makes the argument that program evaluation syntheses should be judged entirely on the theoretical and methodological defensibility of their standards. Yet following these standards leads the WWC to highlight studies that cannot be defended. Examples discussed in my article and in Slavin (2007) include studies of less than 5 hours in duration; studies in which the control group was never taught the content assessed; and a study in which English-proficient students were allowed to help limited English-proficient teammates on the test that was used as the dependent measure in the experimental group, although in the control group the limited English-proficient students did not have such help. Each of these studies was given the WWC's highest rating. As noted by Briggs, the WWC gave Saxon Math its highest rating based on a study that involved two classes and one teacher and used a test keyed to the Saxon content. The WWC gave top ratings to two tutoring programs in which the tutoring was given by research staff, not teachers. It gave a high rating to a math program based on a single study that used only a test of the content focused on by the experimental group but not the control group. The WWC emphasizes all of these studies because they used random assignment, which is valuable because it minimizes one kind of bias. Yet ignoring all other forms of bias, including the obvious bias of the kind apparent in these and many other studies, does a disservice to the readers and to the cause of evidence-based reform. Eventually, we must judge the WWC tree by the WWC fruit.

Inclusiveness

Dynarski argues for "inclusiveness" and cautions against "weeding out" studies that may be "what educators are looking for" (p. 27) He uses this inclusiveness argument to defend the WWC's inclusion of interventions that may last only a few hours or a week, which, he argues, may be just what educators want.

In reality, the WWC excludes the great majority of the studies it identifies, often on trivial grounds explained only in cryptic footnotes in technical appendices. In its middle school math review, it excluded all studies of computer-assisted instruction (CAI) and of instructional process programs (such as cooperative learning), as well as all studies of less than a semester. The WWC elementary math review includes CAI but not instructional process approaches. Its

English-language learner review excludes all studies in which native language was used, even if the outcome measures were in English. To say that educators might be interested in studies of 5 hours' duration but not interested in CAI or cooperative learning or bilingual programs is to strain credulity.

Measures Inherent to Treatments

Briggs raises an important issue that was also central to earlier critiques of the WWC by Schoenfeld (2006) and Confrey (2007). It has to do with the difficult question of measures of achievement outcomes too closely aligned with the treatment group to be fair to the control group. What makes this a thorny question is that there is a continuum of the degree to which measures are slanted toward the content taught in the experimental group. For example, some studies use measures of content seen only by the experimental group, as in studies of a phonemic awareness software program called Daisy Quest. In those studies the control group was not taught phonemic awareness, and the outcome measure included computerized activities similar to the Daisy Quest content, which the control group never saw (e.g., Torgesen et al., 1999). A bit further down the continuum, a study of Everyday Mathematics by Carroll (1998) used an experimenter-made test of a form of geometry emphasized in the experimental group but not the control group. Other studies have used experimenter-made measures that may be fair to the control group, but it is impossible to tell. The WWC completely ignores this issue and includes achievement measures regardless of their link to the experimental curricula, arguing that because one cannot draw a bright line between fair and overaligned measures, it is not necessary to deal with the issue (see Herman, Boruch, Powell, Fleischman, & Maynard, 2006).

There is an argument to be made that measures of experimental treatments should include measures of skills uniquely taught in the experimental group, along with measures of skills taught equally in the experimental and control groups. This argument is often found in studies of mathematics, where a typical finding is that students using an experimental curriculum did no worse than controls on content studied by both but did much better on skills taught only in the experimental classes. This is perhaps appropriately claimed as success by program advocates, who argue that the "extra" learning registered on the experiment-specific test is learning of content that is of particular importance (e.g., advanced problem solving). In reporting on individual studies, the solution is to present both types of outcomes. However, I would argue that only the test of content common to both treatments should be averaged in determining the overall strength of evidence. Otherwise, effects of various treatments are sure to depend more on whether experimenters used measures inherent to treatments than on the actual impact of the program. Furthermore, a policy of valuing treatment-inherent tests could lead evaluators to use *only* such tests and to make them as narrowly associated with the treatment as possible. Someday, someone will teach math in Latin, and then test in Latin.

Small Samples

Dynarski expressed the opinion that I was being contradictory in voicing concern about selection bias but then suggesting the exclusion or downplaying of very small studies. Yet research in medicine and, increasingly, in education finds substantial bias due to small sample size (e.g., Kjaergard, Villumsen, & Gluud, 2001; Slavin, Lake, & Groff, 2007). Small studies need not be excluded, but weighting by sample size, as done by Slavin, Cheung, Groff, and Lake (in press), minimizes the effects of very small studies.

Beyond the potential for bias in small studies, there is an issue of face validity. Educators responsible for whole schools and districts cannot be expected to take seriously research reviews recommending that they implement various "research-proven" programs based on a single small study. Briggs charges that both the WWC and the BEE "appear to conflate sample size . . . with external validity" (p. 21). Yet large sample size is a key indicator of external validity, demonstrating that program effects are found across many settings and in practical circumstances in which experimenters were unlikely to have been able to provide unrealistic levels of support or attention, as they can and do in small studies.

Duration

In criticizing the BEE's requirement of at least 12 weeks' duration for study inclusion, Dynarski notes that educators might well be interested in shorter treatments if they are effective. In fact, the BEE includes brief interventions if they posttest more than 12 weeks after pretest. The reason that the BEE excludes brief experiments that test immediately after the intervention is that in such studies artificial circumstances (such as extra help or intensive focus on the content being taught) could be briefly maintained. In any case, the WWC itself is inconsistent on this point. The WWC math reviews require a minimum duration of a semester, whereas the beginning reading review has no minimum duration.

Levels of Analysis

I would be the first to admit that my arguments on levels of analysis are controversial. In essence, I have argued that the key compromise a program evaluation synthesis should make is to include studies in which treatment or assignment is administered at the group level (e.g., schools, classes) but analyzed at the student level. Chatterji, Sloane, Briggs, and Dynarski all brought up this issue in their comments.

I am well aware that clustered (nested) designs (children within classes or schools) require multilevel analyses, such as hierarchical linear modeling. Yet very few evaluations of educational programs analyze at the proper level. The reason is that it usually requires about 40 schools or classrooms (with a good pretest) to generate adequate statistical power for such analyses (see Raudenbush, 1997). As noted by Sloane and by Dynarski, ignoring clustering overstates statistical significance, although it is unbiased. However, excluding studies that treat at one level and analyze at another would essentially wipe out most of the experimental research evaluating educational programs done over the past 30 years. Other reviews, such as those by the BEE, ignore clustering. Reviews by the Comprehensive School Reform Quality Center, most of those by the EPPI-Centre, and most published meta-analyses ignore the issue. The WWC itself only partially deals with clustering. In cases in which assignment took place at the cluster level, the WWC recomputes statistical significance based on the number of clusters. However, in cases where treatments were applied at the group level but assignment was individual, the WWC does not correct for clustering. For example, imagine an experiment in which 80 children are randomly assigned

to four teachers, two experimental and two control. Random assignment at the individual level solves one aspect of the clustering problem but not the most important one: the fact that children in the same classes are taught by the same teachers and have the same classmates. Almost all of the logic that demands adjustments for clustering based on level of assignment in cluster randomized trials also demands adjustments for clustering when students are randomly assigned to classes but the treatment is delivered at the classroom level, because students within classes cannot be considered independent (and because teachers and classes are confounded with treatment). The absurdity of the WWC policy of treating these two forms of clustering so differently is illustrated by the WWC's treatment of the Williams (1986) study of Saxon Math, in which just 46 students were assigned at random to just two classes, both taught by the same teacher, the study's author. Because assignment was random, this one study trumped four larger matched studies that had near-zero effects. However, the WWC ignored the fact that students in the treatment group were clustered in just two classes. Had the WWC adjusted for clustering (e.g., using a Huber-White procedure), the Williams study would have also been nonsignificant. Moreover, despite its obsessive concern about statistical significance in cluster randomized trials, the WWC accepts effect sizes of +0.25 or more as "potentially positive" anyway, regardless of statistical significance and regardless of sample size, clustering, or other factors (however, nonsignificant studies do not qualify programs for the highest rating).

The net effect of the WWC's clustering adjustment procedures is to make it difficult for class-level or school-level interventions to receive a top rating. Most of the programs that have received top ratings have been one-to-one tutoring approaches, not because tutoring is uniquely effective but because it is relatively easy to randomly assign children to tutors within schools. It is not useful to have a review process that can only tell educators, over and over again, that 1-to-1 instruction is better than 1-to-25, when there is plenty of high-quality research finding positive effects of programs used at the school and classroom levels.

From this time forward, large-scale evaluations should be used to evaluate educational programs and should apply appropriate controls for clustering. Nevertheless, in evaluating the past literature, it is a reasonable compromise to ignore clustering and preserve the information from the hundreds of studies that would otherwise be eliminated.

Conclusion

I am glad that my colleagues took the opportunity to provide thoughtful comments on my article, and I hope that the dialogue carried out in these pages will continue in this and other venues and will draw in additional participants. Program effectiveness syntheses are a key linchpin in the movement toward evidence-based reform, and it is crucial to get them right. I thank the editors of *Educational Researcher* for inviting this interchange, and I thank the participants for their comments.

NOTE

This article was written under funding from the U.S. Department of Education (Grant No. R305AS04008). However, any opinions expressed are those of the author and do not represent positions or policies of the U.S. Department of Education.

REFERENCES

Briggs, D. C. (2008). Synthesizing causal inferences. *Educational Researcher, 37*, 15–22.

Carroll, W. (1998). Geometric knowledge of middle school students in a reform-based mathematics curriculum. *School Science and Mathematics, 98*(4), 188–197.

Chatterji, M. (2008). Synthesizing evidence from impact evaluations in education to inform action. *Educational Researcher, 37*, 23–26.

Confrey, J. (2007). Comparing and contrasting the National Research Council report on evaluating curricular effectiveness with the What Works Clearinghouse approach. *Educational Evaluation and Policy Analysis, 28*(3), 195–213.

Dynarski, M. (2008). Bringing answers to educators: Guiding principles for research syntheses. *Educational Researcher, 37*, 27–29.

Green, J. L., & Skukauskaitė, A. (2008). Becoming critical readers: Issues in transparency, representation, and warranting of claims. *Educational Researcher, 37*, 30–40.

Herman, R., Boruch, R., Powell, R., Fleischman, S., & Maynard, R. (2006). Overcoming the challenges: A response to Alan H. Schoenfeld's "What Doesn't Work." *Educational Researcher, 35*(2), 22–23.

Kjaergard, L. L., Villumsen, J., & Gluud, C. (2001). Reported methodological quality and discrepancies between large and small randomized trials in meta-analyses. *Annals of Internal Medicine, 135*(11), 982–989.

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods, 2,* 173–185.

Schoenfeld, A. H. (2006). What doesn't work: The challenge and failure of the What Works Clearinghouse to conduct meaningful reviews of studies of mathematics curricula. *Educational Researcher, 35*(2), 13–21.

Slavin, R. E. (2007, December 19). The What Works Clearinghouse: Time for a fresh start. *Education Week,* 36–27.

Slavin, R. E. (2008). What works? Issues in synthesizing educational program evaluations. *Educational Researcher*, *37*, 5–14.

Slavin, R. E., Cheung, A., Groff, C., & Lake, C. (in press). Effective programs in middle and high school reading. *Reading Research Quarterly.* Available at www.bestevidence.org

Slavin, R. E., Lake, C., & Groff, C. (2007). *Effective programs in middle and high school mathematics*. Manuscript submitted for publication. Available at www.bestevidence.org

Sloane, F. (2008). Through the looking glass: Experiments, quasi-experiments, and the medical model. *Educational Researcher, 37*, 41–46.

Torgesen, J., Wagner, R., Rashotte, C., Rose, E., Lindamood, P., Conway, T., et al. (1999). Preventing reading failure in young children with phonological processing disabilities: Group and individual responses to instruction. *Journal of Educational Psychology, 91,* 579–593.

Williams, D. D. (1986). *The incremental method of teaching Algebra I.* Research report, University of Missouri, Kansas City.

AUTHOR

ROBERT E. SLAVIN is director of the Center for Research and Reform in Education at Johns Hopkins University, 200 W. Towsontown Boulevard, Baltimore, MD 21204; *rslavin@jhu.edu*. He is also director of the Institute for Effective Education at the University of York, in York, United Kingdom. His research focuses on comprehensive school reform, cooperative learning, research review, and evidence-based reform.

Manuscript received January 14, 2008
Revisions received January 16, 2008
Accepted January 16, 2008