Perspectives on Evidence-Based Research in Education

What Works? Issues in Synthesizing Educational Program Evaluations

Robert E. Slavin

Syntheses of research on educational programs have taken on increasing policy importance. Procedures for performing such syntheses must therefore produce reliable, unbiased, and meaningful information on the strength of evidence behind each program. Because evaluations of any given program are few in number, syntheses of program evaluations must focus on minimizing bias in reviews of each study. This article discusses key issues in the conduct of program evaluation syntheses: requirements for research design, sample size, adjustments for pretest differences, duration, and use of unbiased outcome measures. It also discusses the need to balance factors such as research designs, effect sizes, and numbers of studies in rating the overall strength of evidence supporting each program.

Keywords: evidence-based reform; meta-analysis; research review; What Works Clearinghouse

Throughout the history of education, the adoption of instructional programs and practices has been driven more by ideology, faddism, politics, and marketing than by evidence. For example, educators choose textbooks, computer software, and professional development programs with little regard for the extent of their research support. Evidence of effectiveness of educational programs is often cited to justify decisions already made or opinions already held, but educational program adoption more often follows the pendulum swing of fashion, in which practices become widespread despite limited evidentiary support and then fade away regardless of the findings of evaluations. This situation contrasts with that in fields such as medicine and agriculture, in which the embrace of evidence as a basis for practice has led to dramatic progress, as new and demonstrably more effective practices progressively supplant less effective ones (see Slavin, 1989, 2002).

In recent years, there have been many calls for education to follow other fields in placing far greater reliance on evidence as a basis for adoption of programs and practices (e.g., Borman, 2002; Coalition for Evidence-Based Policy, 2003; Mosteller & Boruch, 2002; Shavelson & Towne, 2002; Slavin, 2002; Towne, Wise, & Winters, 2005). *Evidence-based reform*, the movement toward the use of programs and practices found to be effective in rigorous research, has begun to be advocated in federal policies. For example, the 1997 Obey-Porter Comprehensive School Reform Demonstration (see Slavin, in press) program provided significant funding to help schools adopt "proven, comprehensive schoolwide models." Later, the No Child Left Behind Act (U.S. Department of Education, 2002a) famously recommended use of programs and practices "based on scientifically-based research" more than 100 times. The Institute for Education Sciences (U.S. Department of Education, 2002b) has strongly advocated both expanding research on practical programs using rigorous methods, especially randomized experiments, and using the findings of this research to guide policy and practice.

A key requirement for evidence-based policy is the existence of scientifically valid and readily interpretable syntheses of research on practical, replicable education programs. Educational policy cannot support the adoption of proven programs if there is no agreement on what they are. For this reason, the U.S. Department of Education has sponsored several efforts to synthesize research on educational programs. Its flagship initiative is the What Works Clearinghouse (WWC), but other major initiatives include the Comprehensive School Reform Quality Center (CSRQ) and the Best Evidence Encyclopedia (BEE). The British government has sponsored the Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI-Centre). The international Campbell Collaboration (C2) also sponsors and makes available systematic reviews of research, and of course such syntheses appear in academic journals such as the *Review of Educational Research* (e.g., Borman, Hewes, Overman, & Brown, 2003). Several websites that summarize findings of educational program evaluations have also appeared. These include Social Programs That Work (www.evidencebasedprograms.org) and the Promising Practices Network (www.promisingpractices.net).

The problem is that the methods used in these syntheses vary in fundamental ways, leading to inconsistent conclusions

Educational Researcher, Vol. 37, No. 1, pp. 5–14
DOI: 10.3102/0013189X08314117

regarding which programs and practices have strong evidence of effectiveness. This variation is a potentially serious problem for evidence-based reform, as it could undermine the confidence that educators and policy makers place in the entire enterprise. Academic disagreements are healthy (and inevitable), but it is important to understand the issues, at least, and to agree on basic ground rules for program evaluation syntheses.

The purpose of this article is to discuss key issues in synthesizing research on educational programs, to contrast the methods used in the major synthesis efforts, and to propose solutions to methodological problems inherent in syntheses of program evaluations. The article is intended to help researchers, educators, and other readers of program effectiveness syntheses to understand critical distinctions among synthesis efforts and to be critical readers of this rapidly developing body of reviews.

Major Synthesis Efforts

Although there are many individual syntheses and sources of program evaluation reviews, a few particularly ambitious attempts to synthesize research on many educational programs have produced or are currently producing significant original work. These are briefly described below; more information on them appears throughout this article.

What Works Clearinghouse

The WWC (see http://ies.ed.gov/ncee/wwc/) is the largest of the synthesis efforts. Begun in 2002, the WWC had spent more than $30 million as of 2007. It is currently focusing its reviews in seven areas: beginning reading, elementary math, middle school math, early childhood education, programs for English language learners, dropout prevention, and character education. The contract to manage the WWC was originally awarded to the American Institutes for Research (AIR), but in 2007 the contract was given to Mathematica.

The WWC specifies its inclusion and synthesis procedures in great detail, but in practice it allows considerable variation from one topic area to the next on key issues, such as the minimum study duration required for inclusion. All of the WWC reviews emphasize randomized experiments but include high-quality matched quasi-experiments in a lower category.

The WWC has suffered from an inability to meet its own expectations in terms of completion of reviews. After several false starts and many controversies, the WWC announced in 2004 that several of its key reviews, such as those on beginning reading and middle school math, were about to appear. These and others were not posted until summer 2007 and still have major gaps. Potentially, the WWC is the most important of the synthesis efforts for policy, because it alone carries the endorsement of the U.S. Department of Education.

Best Evidence Encyclopedia

The BEE (see www.bestevidence.org) is a product of the Center for Data-Driven Reform in Education (CDDRE), a U.S. Department of Education–funded research center at Johns Hopkins University. Begun in 2004, CDDRE, whose director is the author of this article, was established to create and evaluate district reform strategies built around the use of proven programs. It initially intended to use the WWC as its source of information on proven programs, but because of the WWC's slow pace, CDDRE researchers created their own set of reviews, using standards and procedures similar to those of the WWC. At this writing, the BEE has completed reviews of elementary math, middle and high school math, middle and high school reading, and reading programs for English language learners. Its website contains links to reviews by the CSRQ on comprehensive school reform and other reviews on several topics. The BEE includes easy-to-read "educator's summaries" of reviews, both those written by CDDRE staff and those written by other reviewers.

Comprehensive School Reform Quality Center

The CSRQ (see www.csrq.org) is at the AIR, the original home of the WWC, but its activity is substantially separate. CSRQ carried out and then updated reviews of research on outcomes of comprehensive elementary and secondary school reform programs (such as Success for All, America's Choice, and Modern Red Schoolhouse) and programs of education service providers (such as Edison Schools). CSRQ used review methods quite different from those of the WWC and of the BEE, emphasizing numbers of studies and statistical significance rather than randomized evaluations and effect sizes. Federal funding for the CSRQ ended in 2007.

Campbell Collaboration

The international C2 (see www.campbellcollaboration.org) is a voluntary organization that prepares and disseminates systematic reviews of existing social science research evidence in education, crime, justice, and social welfare. C2 works to improve the methodology of research synthesis and to disseminate state-of-the-art reviewing methods. Its education reviews evaluate the effectiveness of a range of programs and interventions, such as volunteer tutoring programs and after-school programs, with a strong emphasis on randomized controlled trials.

Evidence for Policy and Practice Information and Co-ordinating Centre

The U.K.-based EPPI-Centre (see www.eppi.ioe.ac.uk), funded primarily by the British government, commissions a wide range of reviews on programs in many areas of education, such as science education, English teaching, and citizenship education. The Department of Children, Schools, and Families funds groups of reviewers to work in each area and allows them to come up with their own standards; thus EPPI's reviews vary widely in breadth, focus, and methodology. Most EPPI education-related reviews focus on variables (e.g., effects of grammar teaching on writing) rather than on specific programs.

Unique Characteristics of Program Evaluation Syntheses

One could argue that methodological and substantive issues in reviews of program evaluations are no different from those in other quantitative syntheses, such as meta-analyses (see, for example, Cooper, 1998; Lipsey & Wilson, 2001). However, there are unique characteristics of program evaluations that should guide the choice of procedures within the meta-analytic canon.

A program is defined here as any set of replicable procedures, materials, professional development, or service configurations that educators could choose to implement to improve student outcomes. A program is distinct from a variable in consisting of a specific, well-specified set of procedures and supports. Class

size, assigning homework, or provision of bilingual education are variables, for example, whereas programs typically are based on particular textbooks, computer software, and/or instructional processes and usually have a name and a specific provider, such as a company, university, or individual.

There are three particularly important characteristics of program evaluation syntheses that should be central to review procedures. First, program evaluation syntheses have high stakes. If evidence-based reform takes hold, the education of millions of children may be affected by these syntheses, and commercial fortunes may be made or lost. It is essential not only that conclusions be correct but also that the process by which they are arrived at be open, consistent, impartial, and in accordance with both science and common sense. Second, the number of studies of most practical programs is very small; if there are any studies at all for a given program, there may be just one. Third, the involvement of commercial companies in program evaluations and in publicizing positive outcomes adds to the possibilities for bias. Publication bias, also known as the "file drawer" problem (the difficulty of finding reports of negative or null evaluations; see Cooper, 1998; Torgerson, 2006) is serious in all quantitative syntheses, but it is heightened for syntheses of program evaluations carried out by companies or their contractors, who have no incentive for or tradition of making negative evaluations available. For example, commercial companies frequently make studies available on their websites or other marketing materials, but these rarely include studies that fail to show positive effects. Studies with positive effects conducted by independent researchers or educators are likely to be sent to the publisher and appear on its website, but other studies may disappear if they are not positive.

These three factors—high stakes, small numbers of studies, and involvement of commercial companies—should lead reviewers to be extremely careful and thorough, reporting in sufficient detail the methodologies, findings, and limitations of each study. In these literatures, flaws cannot be assumed to cancel each other out. Coding for various study characteristics and procedures and then statistically testing to see whether effect sizes correlate with them, as suggested by Abrami and Bernard (2007), Lipsey and Wilson (2001), and others, is rarely possible in program evaluation syntheses because of the small numbers of studies of each program. Computing overall ratings of study quality is also not useful, both because of the small numbers of studies of each program and because there are particular design features that introduce so much bias that they cannot be balanced out by other design features (Juni, Witschi, Bloch, & Egger, 1999). Instead, the reviewer must serve as a detective, looking systematically for studies that provide the best tests of the evidence base for each program. Consistent procedures are essential, but following formulas without attention to the particulars of each study can lead to serious error (see Briggs, 2005).

Minimizing Bias

If there were multiple large-scale, randomized, multiyear evaluations of each of several educational programs, then reviewing the evaluations would be straightforward. Given that this is not the case, however, the reviewer faces a dilemma. One could decide to make inclusion criteria extremely stringent, but the result would be a very small set of programs because few have even a single qualifying study. This is in fact the policy set forth on the Social Programs That Work website, which lists qualifying evidence of achievement effects for only three achievement-focused educational programs: Success for All (Borman et al., 2007; Slavin & Madden 2001) and two tutoring programs, SMART (Baker, Gersten, & Keating, 2000) and Lindamood Phonics (Torgesen et al., 1999).

To include a broader set of studies on a broader set of programs, compromises are needed. The reviewer must decide which compromises are worth making and which are not. Different decisions on this question are what create the differences among synthesis efforts.

In considering standards for review, a useful organizing principle is the need to be strict on issues with potential for bias and liberal on issues that have little such potential. For example, including findings from measures made by the experimenter to assess outcomes taught only in the experimental group has substantial potential for biasing outcomes in favor of the experimental group, so this is an area in which strict definitions should apply. Similarly, failure to control for pretest differences introduces substantial potential for bias, so statistical controls for pretest differences must be a requirement. In contrast, studies that fail to account for clustering (e.g., analyzing at the student level when students were nested within classes or schools) will tend to produce more statistically significant differences than they should, but analysis at the wrong level does not affect individual-level effect sizes and is not biased in one direction or the other (Raudenbush & Bryk, 2002). The WWC sets grade spans for its reviews (e.g., K–3 for beginning reading) and then excludes studies for which data have been collected at these and other grade levels (e.g., K–4) unless they include grade-specific analyses. For example, the WWC excluded several studies of a program called Cooperative Integrated Reading and Composition (CIRC) solely because the studies included Grades 3 to 4 or 2 to 4 and the review was limited to Grades K–3. These large, well-controlled, and (in one case) randomized studies (e.g., Stevens, Madden, Slavin, & Farnish, 1987; Stevens & Slavin, 1995a, 1995b; Stevens, Slavin, & Farnish, 1991), published in the most rigorous journals in education, had more to say to educators than the many small, brief experiments emphasized by the WWC but were rejected on a technicality with little potential to bias outcomes.

In this article, I discuss key decisions faced by reviewers of program evaluations. Table 1 summarizes many of the most important issues and offers suggestions for resolving them. The sections that follow address each issue in detail.

Random Assignment Versus Matching

One of the most contentious issues in syntheses of program evaluations is the role of random assignment. Some of the C2 reviews exclude all studies unless they have used random assignment to treatments. It is impossible for a set of studies to reach the highest categories in the WWC ("meets evidence standards") or the BEE ("strong evidence of effectiveness") without at least one high-quality randomized experiment. In contrast, CSRQ emphasizes the number of statistically significant positive results and does not take random assignment into account.

The importance of random assignment, of course, is that it eliminates initial selection bias (although selection bias can arise after the fact from differential attrition). In a matched study, it

Table 1
Summary of Issues and Suggestions in Program Evaluation Syntheses

Issue	Suggestion
Random assignment vs. matching	Randomized designs should be preferred to matched designs, but large, well-controlled matched designs contribute important information.
Randomized experiments vs. randomized quasi-experiments	Randomized designs with analysis at the unit of assignment should be preferred, but large cluster randomized designs not large enough for hierarchical linear modeling contribute unbiased information.
Matched prospective vs. retrospective quasi-experiments	Among matched studies, prospective studies should be strongly preferred to retrospective comparisons. If there are a sufficient number of higher quality studies, retrospective studies should be excluded.
Sample size	Small studies can have highly variable effects and suffer more from publication bias. They often have confounds with school, teacher, and class effects. Larger studies should be preferred. Weighting by sample size may be used.
Pretest differences	Exclude matched studies in which pretests are not given and those in which pretest differences are more than 50% of a standard deviation. Randomized experiments without pretests are acceptable if attrition is low and equal between experimental and control groups.
Duration	Exclude studies of less than 12 weeks in duration.
Outcome measures	Exclude measures inherent to or potentially biased toward experimental treatments.
Program ratings	Create program ratings according to strength of evidence of effectiveness, balancing median effect size, number of studies, and quality of research design. Strongly emphasize outcomes of large, randomized experiments.

may be that schools or teachers who choose to implement a given program are fundamentally different from other schools or teachers in ways that are not adequately controlled for by pretests or other covariates. The staffs that choose a given treatment might be more highly motivated, reform oriented, or stable than are those in otherwise similar schools that end up in the control group. On the other hand, perhaps schools willing to implement an experimental program are more desperate or less confident in their current programs, and these factors could negatively affect outcomes. Similarly, students assigned to a given program (e.g., gifted, special education) or who volunteer to participate in a given program (e.g., after school, summer school) are likely to differ in ways that controls for pretests and demographics do not fully capture (see Cook, 2001).

In practice, experiments that use random assignment sometimes obtain results different from those obtained in otherwise similar matched studies, and sometimes there is no difference. Heinsman and Shadish (1996) compared randomized and matched studies in four reviews of research on educational interventions and found that in two cases, the two methods led to similar conclusions, whereas in two other cases, they led to somewhat different conclusions. Controlling for pretests and other covariates greatly reduced, but did not eliminate, the differences. Glazerman, Levy, and Myers (2002) also found that use of powerful covariates could greatly reduce but not eliminate differences between randomized and matched studies. This was also the finding of a comparison of randomized and matched studies of dropout prevention programs (Agodini & Dynarski, 2004). However, BEE reviews by Slavin and Lake (in press) and Slavin, Lake, and Groff (2007), using effect size estimates already adjusted for pretests and other covariates in each study, found essentially identical estimates of program effects for randomized and matched experiments. Torgerson (2007) summarized the findings of five meta-analyses of literacy interventions that separately reported effect sizes for randomized and matched studies. Four of the five reported very similar effect sizes for studies using these two designs.

The evidence to date suggests that quasi-experimental studies in which experimental and control groups are well matched, and in which covariates that correlate strongly with pretests (e.g., achievement pretests) are used to adjust outcomes, produce good, if not perfect, estimates of program outcomes, as long as there are no possibilities of selection bias at the individual student level. In other words, among studies comparing one math or reading program with another in which classes receive the treatments, randomized and matched studies may produce similar outcomes; however, in studies of after-school or summer-school programs, or of gifted or special education programs, selection factors are so likely and potentially so consequential that random assignment may be essential. The dropout prevention studies reviewed by Agodini and Dynarski (2004) fall into this latter category. Significantly, the one meta-analysis in Torgerson's (2007) comparison in which effect size estimates differed between randomized and matched studies was a synthesis of one-to-one tutoring for at-risk elementary students by Elbaum, Vaughn, Hughes, and Moody (2000). In such studies, selection bias is likely.

On the other hand, even if random and matched experiments produced very similar outcomes, there are important reasons to prefer randomization. In particular, because of the high-stakes nature of program evaluation syntheses, randomization provides an important safeguard against selection bias. Selection bias may balance out in the long run, over many studies, but in an area in which small numbers of studies determine conclusions about program effects, such balancing cannot be counted on. Random

assignment is essential in building confidence that program outcomes are what they appear to be.

Because of Institute of Education Sciences policies favoring randomized experiments, there are now dozens of experiments in the field, and these show that such studies are feasible (see Borman, 2002; Boruch, 2006; Cook, 2001; Mosteller & Boruch, 2002). In reviews of program evaluations, randomized experiments are justifiably referred to as meeting the gold standard of research design. However, well-matched designs with pretests as covariates can provide good approximations and are often more feasible.

Although randomized experiments should be preferred to matched studies because of the reduction in selection bias inherent in randomized designs, the nature and size of randomized experiments also need to be taken into account in evaluating evidence in a synthesis. First, it is important to be sure that a study claiming random assignment did, in fact, use random assignment. Many researchers consider use of scheduling computers or other procedures under the control of school staff to be "essentially random," but they are mistaken, and numerous such studies report substantial pretest differences despite "random" assignment. Furthermore, many randomized experiments in education are very brief, very artificial, and/or very small and may have serious limitations in both internal and external validity. For example, the Kulik (2003) synthesis of research on computer-assisted instruction (CAI) and the WWC (2007a) beginning-reading topic report both included several studies in which the treatment duration was a few hours. Such brief treatments may be appropriate for laboratory experiments, but they do not inform educators about the likely impact of practical programs. Moreover, they usually create highly artificial conditions (such as one-to-one assistance in studies on technology applications) that could not be maintained over a whole school year.

Issues relating to small, brief, artificial studies are discussed in other sections of this article, but the important point here is that random assignment does not guarantee validity. Entirely appropriate policies promoting experiments using random assignment should not be allowed to lead to an emphasis on studies that are brief, small, artificial, or otherwise of little value to practicing educators.

Randomized Experiments Versus Randomized Quasi-Experiments

Among randomized experiments, those in which teachers, classes, or schools are randomly assigned to treatments are common. The proper analysis for such cluster randomized trials (CRTs) is either hierarchical linear modeling (HLM; Raudenbush, 1997) or analyses of covariance using cluster means. Depending on the effect sizes, correlations between covariates and outcomes, and intraclass correlations, CRTs evaluating educational programs often require 40 or more clusters (schools or classes) for adequate statistical power (Raudenbush, 1997), a practical impossibility for most researchers.

As a result, many researchers assign schools or classrooms at random to treatment and control groups but then analyze at the student level (or use a fixed rather than a random-effects HLM, which can produce similar estimates). Although these procedures are discouraged by methodologists (e.g., Donner & Klar, 2000; Murray, 1998) because they overstate statistical significance, nevertheless their effect sizes are unbiased (Raudenbush & Bryk, 2002) and therefore are of value in syntheses of program evaluations. For example, the WWC corrects all studies with treatments given at the class or school level for clustering, but its technical appendix on this topic states, "Although the point estimates of the intervention's effects based on [studies in which the unit of analysis does not match the unit of assignment] are unbiased, the standard errors of the effect estimates are likely to be underestimated" (WWC, 2007c, p. 12). The value of cluster trials analyzed at the individual level is related to the experiment's number of clusters. If, say, 2 schools are assigned at random to experimental or control treatments, treatment is completely confounded with school, and the results are of less value. If 10 schools are randomly assigned to treatments, however, this is still almost certainly too few for adequate power with HLM, but such a study would nevertheless be valuable because of its lack of bias.

Studies in which schools or classes are randomly assigned to treatments but have too few clusters for multilevel modeling are referred to in the BEE reviews as "randomized quasi-experiments," or RQEs. RQEs are flawed in that they tend to produce more statistically significant positive or negative differences than they should (because analysis at the student level overstates power), but their effect size estimates are unbiased. For this reason, RQEs should be treated as more conclusive than matched studies but less so than true randomized experiments of similar size. Both the WWC and the BEE require at least one randomized experiment with a positive effect for a program to receive the highest rating. However, the BEE allows this experiment to be an RQE; the WWC does not. Instead, the WWC recomputes analyses in RQEs to control for clustering, which almost invariably makes analyses nonsignificant, regardless of effect sizes or student sample sizes.

Matched Prospective Versus Retrospective Quasi-Experiments

Matched studies are not all of one kind. A key design consideration among matched studies is whether the experimental and control groups are designated in advance (a prospective design) or determined after the fact (a retrospective or post hoc design). The distinction between prospective and retrospective designs is of enormous importance in program effectiveness reviews. Retrospective studies may be biased in favor of experimental programs. In comprehensive reviews of research on elementary and secondary math programs, Slavin and Lake (in press) and Slavin et al. (2007) found that retrospective studies had effect sizes almost twice those of prospective matched studies.

In retrospective designs, a group of schools or teachers who have been using a given program, perhaps for many years, is compared after the fact with "control" schools that matched the experimental schools on variables such as pretest achievement scores and demographics (e.g., poverty, race). One problem with such studies is that only the "survivors" are included. Schools that, for example, bought the materials and received the training but abandoned the program before the study took place are not in the final sample, which is therefore limited to more capable or successful schools. For example, Waite (2000) described how 17 schools in a Texas city originally received materials and training for the Everyday Mathematics program. Only 7 schools were still

implementing it at the end of the year, and 6 of those agreed to be in the evaluation. The staffs of the 6 schools may have been more capable or motivated than those of the schools that dropped the program. The comparison group within the same city was likely composed of the full range of more and less capable school staffs, and they presumably had had the same opportunity to implement Everyday Mathematics but chose not to do so. Other post hoc studies, especially those with multiyear implementations, must have also had some number of dropouts, but they typically do not report how many schools took part at first and how many dropped out. The chances are that any school staff able to implement an innovative program for several years is better than staffs that are unable to do so or (even more so) than those that abandoned the program because it was not working. Moreover, schools that see their test scores improving (perhaps for reasons that have nothing to do with the program) are more likely to keep their program than those whose test scores are dropping. As an analog, imagine an evaluation of a diet regimen that studied only people who kept up the diet for a year or more.

Worst of all, retrospective studies usually report outcome data selected from many potential experimental and comparison schools and may therefore report on especially successful schools using the program or on matched control schools that happen to have made particularly small gains, making an experimental group look better by comparison. The fact that researchers in retrospective studies often have pre- and posttest data from state test scores readily available on hundreds of potential matches, and may deliberately or inadvertently select the schools that show the program to best effect, means that readers must take results from after-the-fact comparisons with a grain of salt.

Despite all of these concerns, retrospective studies are included in the WWC and the BEE for one reason: Without them, there would be no evidence at all concerning most of the commercial textbook series used by the vast majority of schools. As long as the experimental and control groups are well matched at pretest on achievement and demographic variables and meet other inclusion requirements, they may be included, with appropriate caveats. However, when the field matures enough to have many randomized and prospective matched studies available, this category should be excluded.

Sample Size

Many studies of educational programs use very small samples. Small numbers of students create obvious problems of inadequate statistical power, but small numbers of classes and schools create additional problems of confounding. As noted earlier, in a study in which children are randomly assigned to Teacher A teaching the experimental treatment or Teacher B teaching the control class, treatment effects are completely confounded with teacher and class effects. The larger the number of independent units in each treatment group, the less confounding there is.

Small studies are likely to be biased in favor of the experimental group because small studies with null or negative results are more likely to be impossible to find than are otherwise similar large studies. As noted earlier, studies with small sample sizes tend to have more extreme effect sizes, both positive and negative, especially because factors such as school, teacher, and class effects can greatly affect outcomes in small studies but tend to even out in larger studies (Givens, Smith, & Tweedie, 1997). Small studies with zero or negative effects are less likely to be published or reported in any form than are larger studies with zero or negative effects (Sterne, Gavaghan, & Egger, 2000). Because of their cost and difficulty, the results of large studies are likely to be available, at least in technical reports or dissertations, regardless of their findings. In meta-analyses that synthesize many studies, a procedure called "trim and fill" (Taylor & Tweedie, 1998) is sometimes used to estimate the number of presumed missing small studies with negative or null outcomes to balance against the excessive estimates from the small studies with positive effects that were therefore published. Other statistical procedures to detect and control for publication bias have also been described (e.g., Dear & Begg, 1992; Givens et al., 1997; Hedges, 1992; Rothstein, Sutton, & Borenstein, 2005). However, these procedures are rarely used and are not practical with the small numbers of studies likely to exist for any given program in program evaluation syntheses.

Small studies may allow researchers to spend a great deal of time ensuring exemplary implementation of experimental treatments, but doing so is difficult in large studies, which are more likely to simulate the realistic conditions that the treatment will face when it is scaled up and used as a routine part of schools' curricula. Cronbach et al. (1980) warned against taking too seriously the results of small studies that evaluate "superrealizations"—ideal, nonreplicable implementations of experimental treatments.

In practice, sample size can make a substantial difference in effect size. In the Slavin and Lake (in press) BEE review of elementary mathematics, the median effect size for qualifying CAI studies with sample sizes of fewer than 250 students was +0.21, whereas the median for studies with larger sample sizes was only +0.11. In the Slavin et al. (2007) review of secondary math programs, the median effect size for CAI studies with sample sizes of fewer than 250 was +0.21; the median for studies with larger sample sizes was only +0.07. Corresponding median effect sizes were +0.53 for three small studies of CAI in secondary reading and +0.18 for seven larger studies (Slavin, Cheung, Groff, & Lake, in press). Similar patterns were seen for all types of interventions in all three BEE reviews. It is important to note that small studies are not inherently biased, but collectivities of small studies tend to be biased because of file drawer effects and other problems.

Unfortunately, random assignment studies tend to have small sample sizes, especially when individual students are assigned at random. And because of confounding with teacher and school effects as well as publication bias, these small randomized studies tend to be biased toward positive outcomes (Givens et al., 1997; Sterne et al., 2000). A large, prospective matched study may provide more meaningful and reliable information than a small, randomized one. Limiting reviews to randomized experiments may inadvertently introduce bias if most randomized studies are small. For all of these reasons, large studies, especially those that use random assignment to conditions, should be strongly emphasized in program evaluation reviews. The WWC excludes studies in which there is only one teacher or school in each condition but otherwise does not attend to sample size. The BEE, in contrast, strongly emphasizes evaluations with more than 250 students in 10 classes or schools. Smaller studies are not excluded, but they

are downplayed in outcome summaries unless sample sizes across multiple small studies collectively reach 250 students. The sample size problem might also be solved by weighting, and this was done by Borman et al. (2003) in their review of studies of comprehensive school reform programs.

The procedures used by the WWC leads to a situation in which very small (but randomized) experiments largely determine the ratings given to many programs. For example, the WWC (2007b) gave its top rating, "positive effects," to a middle school math program, Saxon Math. The randomized study that qualified Saxon Math for this rating was an unpublished report by Williams (1986) involving 46 students taught by one teacher. The only outcome measure was a test made up by Williams himself that was closely aligned with the Saxon Math curriculum (but not the curriculum used in the control group). The effect size reported by the WWC for this study was +0.65, yet four other qualifying studies that used conventional measures had a median effect size of only +0.06. Because the Williams study used random assignment, however, its very positive outcome trumped the others. Similarly, the only program to receive a positive-effects rating in the English language learners topic report was one called Peer Tutoring and Response Groups. This program qualified on the basis of a 4-week study by Prater and Bermudez (1993) of 46 children in which children in the experimental group were able to work with English-proficient group mates on the composition from which the outcome measure was computed, whereas the control students worked alone. The effect size across four measures of composition (not English language proficiency) was +0.46.

What these and many other examples illustrate is that a focus on randomized studies without attention to sample size and other design elements that also have potential to introduce bias can lead to illogical conclusions.

Pretest Differences

In studies of academic achievement, pretests and other factors (such as demographics) are almost always powerful predictors of posttests. Statistical controls for pretest differences, such as analyses of covariance (ANCOVA), regressions, or HLM controlling for pretests, work well when experimental–control pretest differences are small. However, large pretest differences cannot be adequately controlled for, as the underlying distributions may be very different, especially when ceiling or floor effects are possible. ANCOVAs or other statistical controls will tend to undercontrol or (less often) overcontrol (Shadish, Cook, & Campbell, 2002). Use of propensity matching or similar procedures may reduce the problem of comparing similar students in dissimilar groups (Dehejia & Wahba, 1999), but this procedure is uncommon in program evaluations in education. When pretest differences are greater than a half standard deviation, studies should be excluded.

Posttest effect sizes should always be adjusted for pretest differences, whether or not they are significant. Ideally, posttests should be statistically adjusted for pretests and other covariates, but if adjusted posttests are not available, pretest effect sizes should be subtracted from posttest effect sizes. Only in true randomized experiments with minimal attrition should unadjusted posttest means be used, and even in such studies, posttests should be adjusted for pretests if the pretests are available. Nonrandomized studies lacking pretests or other highly correlated variables indicating initial equivalence should be excluded, and the WWC and the BEE do so.

Duration

Educators and policy makers considering research on educational programs need to be sure that the evidence they are shown relates to practical programs that can be used over extended time periods, not theoretically interesting but impractical procedures that could never be replicated for extended periods. For example, an early WWC review on peer-assisted learning in elementary schools (later removed) included numerous studies of a few hours in duration in which neither experimental nor control groups received any teaching. In its beginning-reading review, the WWC included and gave its highest rating to phonemic awareness software called Daisy Quest, which was evaluated in studies of less than 5 hours. In the Daisy Quest studies, members of the research team sat with small groups of students as they worked on the computers, providing assistance that clearly could not have been provided in a longer study. Similarly, in an 8-week study of a tutoring program called SpellRead, project personnel were used as tutors (Rashotte, MacPhee, & Torgesen, 2001), yet the program was highly rated by the WWC.

In general, brief studies are low in external validity. For this reason, various program evaluation reviews set minimum durations for inclusion of studies. Different WWC reviews use different duration criteria, from none at all (in beginning reading) to a semester (in elementary math). The BEE uses a 12-week criterion.

Outcome Measures Inherent to Treatments

A difficult issue in reviews of program evaluations relates to studies in which outcome measures assess skills taught in the experimental group but not the control group. As noted earlier, measures inherent to the experimental treatment have substantial potential to bias findings toward positive effects. This was a serious problem in the Williams (1986) study of Saxon Math and the Prater and Bermudez (1993) study of Peer Tutoring and Response Groups, cited earlier. An extreme example is the series of brief (5 hours or less) and small (69 students or fewer) studies evaluating Daisy Quest, a computerized program used to teach phonemic awareness in Grades K through 1, which received the highest rating (positive effects) in the WWC reviews of both beginning reading programs and early childhood programs. In the studies (e.g., Barker & Torgesen, 1995), the control groups were not taught phonemic awareness at all. Worse, some of the outcome measures were activities from Daisy Quest, which the control group had, of course, never seen. One of the studies, by Foster, Erickson, Foster, Brinkman, and Torgesen (1994), compared children taught phonemic awareness by a teacher to those taught using Daisy Quest on the computer. Those taught by the teacher did much better than those taught with Daisy Quest, although both groups, not surprisingly, performed much better than children who were not taught any phonemic awareness. Daisy Quest received the highest possible rating from the WWC for its effects on "alphabetics" because of its use of random assignment.

Another previously cited example, also from the WWC, involves a study by Carroll (1998) of Everyday Mathematics. The

only outcome measure was an assessment of a form of geometry taught in Everyday Mathematics but not in the control group.

Outcome measures focusing on content taught in experimental groups but not control groups should not be included in syntheses of program evaluations, as they unfairly favor the experimental treatments. Numerous studies (e.g., Crawford & Snider, 2000; Van Dusen & Worthen, 1994; Ysseldyke et al., 2003) have used both national standardized tests and developer-made tests, and effect sizes are invariably much more positive on the latter measures. The developer-made tests are by definition intended to assess outcomes taught in the program, and such tests are unfair to students exposed to different content. When developers have a good rationale to assert that the content taught and assessed in their program is more valuable than the content assessed on standardized or other neutral assessments, there is nothing wrong with pointing out effects on such measures; however, these outcomes should not be included in comparative reviews of research on alternative programs, because doing so skews the review in favor of programs that use developer-made assessments and against programs evaluated on the types of measures for which students and schools are held accountable. For this reason, measures inherent to the experimental treatment are excluded in the BEE reviews.

Program Ratings

In a program evaluation synthesis, readers ultimately want an easily interpreted, well-justified rating of the strength of the evidence base and the size of the anticipated effects for each program. Reviewers may be uncomfortable with this, knowing the complexity and uncertainties behind their conclusions; the WWC, for example, states that "the WWC does not endorse any interventions," and the CSRQ reports have similar language. Yet readers are sure to interpret ratings as endorsements of the research base, if not of the program itself. For this reason, the program rating process must be taken very seriously.

The rating process is more complex than it looks, and different program evaluation syntheses have used very different methods. The problem is that several attributes of a body of studies must be balanced.

1. *Effect size.* A set of experiments could be summarized in terms of a mean or median effect size, perhaps by doing a miniature meta-analysis for each program. This approach can provide a common metric for all programs to easily express differences between experimental and control groups in percentile ranks. The WWC, for example, reports the experimental group's advantage in percentile ranks represented by a given effect size, setting the control group at 50.

   The problem with reporting average effect sizes is that they can be misleading if the number of studies is small, especially if the studies themselves are small or are otherwise flawed. A mean effect size does not indicate the degree of confidence behind the number. In principle, a single, small, flawed study could give an inflated effect size that would look much more positive than the evidence from dozens of large, high-quality studies.
2. *Statistical significance.* Statistical significance of positive or negative outcomes can be used as an important factor in characterizing outcomes, but this has many problems as well. Emphasizing statistical significance tends to favor large studies, even those with very small effect sizes; for example, an enormous study of National Science Foundation–supported math curricula found significant differences with effect sizes as small as +0.06 (Sconiers, Isaacs, Higgins, McBride, & Kelso, 2003). Furthermore, it is unclear what to do when some outcome measures are significant and some are not.
3. *Number of studies.* A program supported by a large number of studies finding positive effects has stronger evidence than one with few studies, but emphasizing the number of studies can lead to emphasizing programs that happen to have a large number of small, potentially flawed studies or small effects. The CSRQ reviews place the strongest emphasis on numbers of studies in combination with statistical significance, requiring that a program have at least 10 qualifying studies and at least 75% of comparisons statistically significant and positive to be placed in the highest category.
4. *Research design.* Ideally, the studies that determine program ratings should use random assignment to treatments. Some of the C2 reviews required random assignment as an inclusion criterion. However, randomized studies are few in number, and many are very small, very brief, very artificial, and/or very old. Given the increasingly common finding that in studies in education, randomized and well-matched studies tend to produce similar effect sizes (see Torgerson, 2007), the rationale for restricting attention to randomized studies alone is diminished.

All program evaluation syntheses that use ratings try to balance some or all of these factors, but to varying degrees. To receive WWC's highest rating, positive effects, a program must have at least one study that used random assignment and had significant positive effects and at least one additional positive study that met WWC's "meet evidence standards with reservations" standard; moreover, there must be no studies of the program that found significant negative effects. To receive the BEE's highest rating, "strong evidence of effectiveness," requires at least one large randomized study ($N > 250$) or multiple small studies with a collective sample size of 250, a second large randomized or matched study, and a median effect size of at least +0.20. In both WWC and BEE syntheses, however, programs can qualify for a second rating category with high-quality matched studies. Borman et al. (2003) balanced mean effect sizes and numbers of studies in their categorization of comprehensive school reform programs.

As long as relatively stringent inclusion criteria have already been applied to the original studies—to weed out those with poor matches, poor controls for pretest differences, very small sample sizes, brief durations, and measures slanted toward the treatment groups—then it may not matter as much which pooling strategy is used. The danger is that if poor studies are not excluded, either a single study with an anomalously large effect size or a set of studies with a consistent bias will influence final ratings. In that event, the legitimacy of the entire enterprise would be undermined.

Conclusion

Evidence-based reform has the potential to substantially change the practice of education and to make education research far more central to education policy. Practitioner-friendly syntheses of research on practical programs play an essential role in establishing the idea that there is evidence worth paying attention to. It is of great importance to make such reviews as valid, unbiased, and meaningful as possible for their intended purpose. It is also important that researchers and educators understand the critical issues behind the various program effectiveness reviews so that they can intelligently interpret their conclusions.

I hope that this article will be one of many discussing the issues that need to be considered in syntheses of program evaluation research. Clear, thoughtful syntheses in many areas are crucial to providing practitioners, policy makers, and researchers with valid information that they can use with confidence to address the real problems of educating all children.

NOTE

This article was written under funding from the Institute of Education Sciences, U.S. Department of Education (Grant No. R305A040082). However, any opinions expressed are those of the author and do not necessarily represent positions or policies of the institute. I would like to thank Harris Cooper, Carole Torgerson, Steven Ross, Bette Chambers, Alan Cheung, Philip Abrami, Marlene Darwin, Jon Baron, Mark Newman, and anonymous reviewers for comments on an earlier draft.

REFERENCES

Abrami, P. C., & Bernard, R. M. (2007). *Statistical control vs. classification of study in meta-analysis.* Manuscript submitted for publication.

Agodini, R., & Dynarski, M. (2004). Are experiments the only option? A look at dropout prevention programs. *Review of Economics and Statistics, 86*(1), 180–194.

Baker, S., Gersten, R., & Keating, T. (2000). When less may be more: A 2-year longitudinal evaluation of a volunteer tutoring program requiring minimal training. *Reading Research Quarterly, 35*(4), 358–366.

Barker, T. A., & Torgesen, J. K. (1995). An evaluation of computer-assisted instruction in phonological awareness with below average readers. *Journal of Educational Computing Research, 13*(1), 89–103.

Borman, G. D. (2002). Experiments for educational evaluation and improvement. *Peabody Journal of Education, 77*(4), 7–27.

Borman, G. D., Hewes, G. M., Overman, L. T., & Brown, S. (2003). Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research, 73*(2), 125–230. Available from http://www.bestevidence.org

Borman, G. D., Slavin, R. E., Cheung, A., Chamberlain, A., Madden, N. A., & Chambers, B. (2007). Final reading outcomes of the national randomized field trial of Success for All. *American Educational Research Journal, 44*(3), 701–703.

Boruch, R. (2006, April). *Ethical standards, evidence standards, and randomized trials: Error flees but slowly.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Briggs, D. C. (2005). Meta-analysis: A case study. *Evaluation Review, 29*(2), 87–127.

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

Coalition for Evidence-Based Policy. (2003). *Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide.* Washington, DC: U.S. Department of Education.

Cook, T. D. (2001). *Reappraising the arguments against randomized experiments in education: An analysis of the culture of evaluation in American schools of education.* Unpublished manuscript, Northwestern University.

Cooper, H. (1998). *Synthesizing research* (3rd ed.). Thousand Oaks, CA: Sage.

Crawford, D. B., & Snider, V. E. (2000). Effective mathematics instruction: The importance of curriculum. *Education and Treatment of Children, 23*(2), 122–142.

Cronbach, L. J., Ambron, S. R., Dornbusch, S. M., Hess, R. O., Hornik, R. C., Phillips, D. C., et al. (1980). *Toward reform of program evaluation: Aims, methods, and institutional arrangements.* San Francisco: Jossey-Bass.

Dear, K., & Begg, C. (1992). An approach for assessing publication bias prior to performing a meta-analysis. *Statistical Science, 7*(2), 237–245.

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evolution of training programs. *Journal of the American Statistical Association, 94,* 1053–1062.

Donner, A., & Klar, N. (2000). *Design and analysis of cluster randomized trials in health research.* New York: Oxford University Press.

Elbaum, B., Vaughn, S., Hughes, M. T., & Moody, S. W. (2000). How effective are one-to-one tutoring programs in reading for elementary students at risk for reading failure? A meta-analysis of the intervention research. *Journal of Educational Psychology, 92*(4), 605–619.

Foster, K., Erickson, G., Foster, D., Brinkman, D., & Torgesen, J. (1994). Computer administered instruction in phonological awareness: Evaluation of the Daisyquest program. *Journal of Research and Development in Education, 27*(2), 126–137.

Givens, G., Smith, D., & Tweedie, R. (1997). Publication bias in meta-analysis: A Bayesian data-augmentation approach to account for issues exemplified in the passive smoking debate. *Statistical Science, 12*(4), 221–250.

Glazerman, S., Levy, D. M., & Myers, D. (2002). *Nonexperimental replications of social experiments: A systematic review.* Washington, DC: Corporation for the Advancement of Policy Evaluation.

Hedges, L. (1992). Modeling publication selection effects in meta-analysis. *Statistical Science, 7*(2), 246–255.

Heinsman, D. T., & Shadish, W. R. (1996). Assignment methods in experimentation: When do nonrandomized experiments approximate answers from randomized experiments? *Psychological Methods, 1*(2), 154–169.

Juni, P., Witschi, A., Bloch, R., & Egger, M. (1999). The hazards of scoring the quality of clinical trials for meta-analysis. *Journal of the American Medical Association, 282*(11), 1054–1060.

Kulik, J. A. (2003). *Effects of using instructional technology in elementary and secondary schools: What controlled evaluation studies say* (SRI Project Number P10446.001). Arlington, VA: SRI International. Available from http://www.bestevidence.org

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis.* Thousand Oaks, CA: Sage.

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

Murray, D. M. (1998). *Design and analysis of group-randomized trials.* New York: Oxford University Press.

Prater, D. L. & Bermudez, A. B. (1993). Using peer response groups with limited English proficient writers. *Bilingual Research Journal, 17*(1/ 2), 99–116.

Rashotte, C. A., MacPhee, K., & Torgesen, J. K. (2001). The effectiveness of a group reading instruction program with poor readers in multiple grades. *Learning Disabilities Quarterly, 24*(2), 119–134.

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

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods.* Thousand Oaks, CA: Sage.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.). (2005). *Publication bias in meta-analysis: Prevention, assessment, and adjustments.* Chichester, UK: John Wiley.

Sconiers, S., Isaacs, A., Higgins, T., McBride, J., & Kelso, C. (2003). *The Arc Center tri-state student achievement study.* Lexington, MA: COMAP.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002) *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton-Mifflin.

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. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher, 31*(7), 15–21.

Slavin, R. E. (in press). Comprehensive school reform. In T. Good (Ed.), *21st century education: A reference handbook.* Thousand Oaks, CA: Sage.

Slavin, R. E., Cheung, A., Groff, C., & Lake, C. (in press). Effective reading programs for middle and high schools: A best-evidence synthesis. *Reading Research Quarterly.*

Slavin, R. E., & Lake, C. (in press). Effective programs in elementary mathematics. *Review of Educational Research.* Available from http://www.bestevidence.org

Slavin, R. E., Lake, C., & Groff, C. (2007). *Effective programs in middle and high school mathematics.* Baltimore: Center for Data-Driven Reform in Education, Johns Hopkins University. Manuscript submitted for publication. Available from http://www.bestevidence.org

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

Sterne J., Gavaghan, D., & Egger, M. (2000). Publication and related bias in meta-analysis: Power of statistical tests and prevalence in literature. *Journal of Clinical Epidemiology, 53,* 1119–1129.

Stevens, R. J., Madden, N. A., Slavin, R. E., & Farnish, A. M. (1987). Cooperative Integrated Reading and Composition: Two field experiments. *Reading Research Quarterly, 22*(4), 433–454.

Stevens, R. J., & Slavin, R. E. (1995a). The cooperative elementary school: Effects on students' achievement, attitudes, and social relations. *American Educational Research Journal, 32*(2), 321–351.

Stevens, R. J., & Slavin, R. E. (1995b). Effects of cooperative learning approach in reading and writing on academically handicapped and nonhandicapped students. *Elementary School Journal, 95*(3), 241–262.

Stevens, R. J., Slavin, R. E., & Farnish, A. M. (1991). The effects of cooperative learning and direct instruction in reading comprehension strategies on main idea identification. *Journal of Educational Psychology, 83*(1), 8–16.

Taylor, S., & Tweedie, R. (1998). *A non-parametric "trim and fill" method of assessing publication bias in meta-analysis.* Denver: University of Colorado Health Sciences Center.

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.

Torgerson, C. J. (2006). Publication bias: The Achilles' heel of systematic reviews? *British Journal of Educational Studies, 54*(1), 89–102.

Torgerson, C. J. (2007). The quality of systematic reviews of effectiveness in literacy learning in English: A "tertiary" review. *Journal of Research in Reading, 32*(3), 287–315.

Towne, L., Wise, L. L., & Winters, T. M. (Eds.). (2005). *Advancing scientific research in education.* Washington, DC: National Academies Press. Available from http://www.NAP.edu

U.S. Department of Education. (2002a). *No Child Left Behind: A desktop reference.* Washington, DC: Author. Available from http://www.ed.gov/offices/OESE/reference

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

Van Dusen, L., & Worthen, B. (1994). The impact of integrated learning system implementation on student outcomes: Implications for research and evaluation. *International Journal of Educational Research, 21,* 13–24.

Waite, R. E. (2000). *A study on the effects of Everyday Mathematics on student achievement of third-, fourth-, and fifth-grade students in a large North Texas urban school district.* Unpublished doctoral dissertation, University of North Texas.

What Works Clearinghouse. (2007a). *Beginning reading topic report.* Retrieved December 19, 2007, from http://ies.ed.gov/ncee/wwc/reports/beginning_reading/

What Works Clearinghouse. (2007b). *Intervention report, Saxon Middle School Math, technical appendix.* Retrieved December 19, 2007, from http://ies.ed.gov/ncee/wwc/pdf/techappendix03_17.pdf

What Works Clearinghouse. (2007c). *Technical details of WWC-conducted computations.* Retrieved December 19, 2007, from http://ies.ed.gov/ncee/wwc/pdf/conducted_computations/pdf

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

Ysseldyke, J., Spicuzza, R., Kosciolek, S., Teelucksingh, E., Boys, C., & Lemkuil, A. (2003). Using a curriculum-based instructional management system to enhance math achievement in urban schools. *Journal of Education for Students Placed at Risk, 8*(20), 247–265.

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 16, 2007
Revisions received July 12, 2007, and September 21, 2007
Accepted October 23, 2007