



Comments on Slavin

Bringing Answers to Educators: Guiding Principles for Research Syntheses

Mark Dynarski

Research syntheses are appealing because they enable decision makers to determine quickly whether policies, programs, and practices will have effects on student achievement and, if so, the magnitudes of the likely effects. Such syntheses should present objective, clear, scientifically accurate, and defensible evidence in terms that educators can use in making their decisions. The syntheses should also give the most weight to research with strong causal validity and should consider sources of bias and use procedures that reduce or eliminate bias. To that end, it is hoped that theoretical development of standards will help researchers to assess the trade-offs that need to be made in weighing the evidence to be used in syntheses. Creating methods for assessing the extent of evidence in syntheses will enhance their usefulness in evidence-informed decision making.

Keywords: causal inference; program evaluation; research syntheses

The success of the Cochrane Collaboration in health research points to the potential value of research syntheses in education. The use of evidence in education decision making is likely to grow, with continued impetus from the No Child Left Behind Act (2001) and other accountability demands.

Decision makers use research syntheses to determine quickly whether policies, programs, and practices will have effects on student achievement and, if so, the magnitudes of the likely effects. Such syntheses should present evidence that is objective, clear, scientifically accurate, and defensible, and they should be stated in terms that educators can use in making their decisions. These features will help educators to make evidence-informed decisions that can improve student achievement.

In “What Works: Issues in Synthesizing Educational Program Evaluations” (this issue of *Educational Researcher*, pp. 5–14), Robert E. Slavin raises questions on synthesizing program evaluations that I confront in my current work as director of the What Works Clearinghouse (WWC) and in carrying out experimental and quasi-experimental education evaluations. The variation in approaches to conducting evaluations makes the effort to synthesize findings a challenge. I agree with Slavin that research

syntheses and ratings of research evidence should be anchored in a set of principles that clearly indicate how trade-offs in combining research findings are made. That evaluations of a given program are often few in number and commercial stakes often high does not seem to me to be a basis for care in the development and application of ratings, as Slavin argues. I think ratings should be developed and applied carefully regardless of the number of program evaluations and the financial stakes involved. Keeping in mind who the audience is, what kinds of information from syntheses can best help that audience, and how to present the information so that it is understood and acted on seem central to the effectiveness of syntheses.

The audience of educators is vast and diverse. The United States alone has more than fifteen thousand school districts, more than a hundred thousand schools, hundreds of thousands of administrators and board members, millions of teachers, and a vast network of technical assistance providers, advocacy and policy organizations, and curriculum and product developers. Decision makers within this vast enterprise have different objectives and face different resource limits in trying to meet their objectives. With different goals and resource limits, decision makers use all kinds of research evidence to decide which curricula and interventions to use in their schools; and these decision makers, like investors of any kind, usually are averse to risk. If two courses of action are viewed as having the same outcome, on average, but different variability around the average outcome, decision makers will prefer the course of action with the smaller variability.

In settings where educators have to make many decisions and are confronted with reams of research generated by entities with financial investments at stake, the inclusiveness and independence of the reviews underlying research syntheses should be a central concern. Selective exclusion of research requires great caution, as selectivity can be interpreted as compromising scientific objectivity for purposes that educators cannot discern and may misinterpret.

Inclusive syntheses may also be useful in providing what educators are looking for. For example, urban school districts that are under pressure to raise reading scores may want to know about the effectiveness of reading interventions that are inexpensive and straightforward to implement. Some of these interventions may last only a few hours or weeks. Arguing that syntheses should screen out these interventions is tantamount to arguing that

researchers should give educators information only about what the researchers believe are appropriate interventions.

Syntheses Should Summarize All Evidence

Slavin tackles the issue of considering which standards are appropriate. For example, he approvingly notes that research syntheses should give greater weight to experiments because they are less subject to selection bias. He does not develop a theoretical framework enabling the properties of various standards to be assessed, and ultimately the argument devolves to noting in various ways that results of syntheses will vary when standards vary.

Without a theory, arguing for the superiority or necessity of various standards has a rhetorical feeling. The strategy implicitly used in Slavin's article is to judge standards by the results of applying them, such as by looking at which interventions end with which ratings. This approach is ad hoc without a higher set of standards. In the article, Slavin proposes to use scientific principles and "common sense" to judge standards. Relying on scientific principles argues for a theoretical framework. I interpret using common sense to mean appealing to broad agreement in the research community about what should be observed when evidence is rated; but if such broad agreement exists, the point of the article is lost.

Slavin argues that syntheses can arrive at "illogical conclusions" when they focus on some aspects of study design (such as whether they are experiments or quasi-experiments) and not others (such as sample size and intervention duration). But if the standards for a synthesis are employed correctly, the conclusions are logical because the procedures are deductive. For example, the WWC reviews the various studies of an intervention and then applies a scheme that arrives at a rating such as "positive" or "no discernible effects," or the like. How the WWC reviews the studies and arrives at the rating is straightforward and replicable. That some might object to the rating assigned to an intervention does not mean that the conclusions underlying the rating are illogical.

The lack of a set of theoretical principles and the need to fall back on common sense perhaps explain some of the contradictions in the article. For example, Slavin argues for "weeding out" studies, such as those of short-duration interventions and those with small samples. But it is contradictory to argue both that using experimental designs to reduce selection bias is desirable and that excluding some interventions and studies, which also creates a kind of selection bias, is desirable. Resolving this contradiction requires making a case that one type of selection bias (sample members choosing to participate in an intervention) is more deleterious in its effects on evidence than the other type of selection bias (researchers deciding which interventions and studies educators should learn about). The article does not make this case. Certainly it is possible that the findings from some studies are due to publication bias or arise from local conditions that are unusual or hard to replicate. But if syntheses review all the evidence and apply sound standards, educators can make up their own minds about whether the findings are credible or whether the implementation conditions are unrealistic and not useful to them.

Another contradiction arises in Slavin's discussion of bias and how it affects rating schemes. Slavin argues that researchers synthesizing studies should be concerned about issues with potential

for bias and less concerned about issues for which there is little potential for bias. Presenting unbiased information about effects is an attractive feature in a research synthesis. But Slavin submits that cluster-level variance is an issue that has little potential for bias. This point is of more than theoretical interest. Many educational evaluations are designed with clusters (classrooms or schools) and do not adjust their estimates of variance for this feature, thereby overstating statistical significance. Because the WWC rating scheme considers whether estimates are statistically significant, underestimating variances would lead to too many interventions' being rated as having positive or potentially positive effects.

Two points can be made here. First, Slavin is using an unconventional definition of bias. In the statistical sense, an estimator is defined as biased if its expected value does not equal the true parameter. The definition applies to estimators of variance just as much as to estimators of means. Not adjusting for cluster variance induces a possibly large degree of bias in estimating variances; it is unclear why syntheses should not be concerned about this kind of bias. Second, it is contradictory for Slavin to encourage researchers to view schemes for rating research as flawed if they are sensitive to small studies or ones of short-duration interventions, while on the other hand encouraging researchers to downplay correct estimates of statistical significance in rating research. It is unclear why a researcher conducting a synthesis would choose to rate too many interventions as having evidence of positive effects rather than apply a straightforward correction to variance estimates.

Assessing the Extent of Evidence Will Be Useful

A theme that Slavin touches on at various points relates to what I will label "the extent of evidence." In assessing the potential efficacy of an intervention, decision makers are well served when interventions have been tested in a broad range of circumstances and local contexts. Here is where the risk-averse nature of the decision maker comes into play. A decision maker faces a higher degree of risk when interventions are tested in only a few situations or only in those that seem to be far removed from the decision maker's context. In contrast, knowing that a particular intervention under consideration has shown large positive effects in a range of situations and contexts in which it has been tested is powerful and desirable information because it suggests the risks of implementing it are moderate.

Developing a scheme that characterizes the extent of evidence is an important challenge for researchers to tackle in the future. In theory, the extent of evidence is affected by the number of studies, the number of units in the studies (such as districts, schools, teachers, and students), and the internal and external validity of the studies (including whether the studies are experiments and whether the units are nationally representative). The WWC rates the extent of evidence on the basis of a combination of the number of studies and the total sample size of the combined studies. The Best Evidence Encyclopedia rates a study as large if it has more than 10 classrooms or 250 students. Helping educators factor the extent of evidence into their decisions requires more information than either of these sources currently provides about the diversity of settings, local contexts, and counterfactuals.

Looking Ahead

The WWC recently began releasing practice guides based on research findings that present concrete recommendations to educators. The interest in these guides is a reminder that educators perhaps are less interested in the routes taken to arrive at answers than in the answers themselves. Nonetheless, more conceptual thinking about standards will be valuable, and syntheses should be a matter for continued discussion within the broader research community whose efforts are being synthesized.

NOTE

The views expressed here are those of the author and do not necessarily represent the views of Mathematica Policy Research, Inc., or the U.S. Department of Education.

REFERENCES

No Child Left Behind Act of 2001, Pub. L. No. 107-110 (2001).
Slavin, R. E. (2008). What works? Issues in synthesizing educational program evaluations. *Educational Researcher*, 37, 5–14.

AUTHOR

MARK DYNARSKI is a senior fellow and an associate director at Mathematica Policy Research, Inc., 600 Alexander Park, Princeton, NJ 08540; *MDynarski@mathematica-mpr.com*. He is also the director of the What Works Clearinghouse. His research focuses on education program evaluations and research methodology.

Manuscript received January 13, 2008

Accepted January 14, 2008