



Comments on Slavin

Synthesizing Causal Inferences

Derek C. Briggs

When causal inferences are to be synthesized across multiple studies, efforts to establish the magnitude of a causal effect should be balanced by an effort to evaluate the generalizability of the effect. The evaluation of generalizability depends on two factors that are given little attention in current syntheses: construct validity and external validity. The author concludes with six concrete suggestions for ways that future syntheses could be improved to make them more relevant for educational policy makers.

Keywords: causal inference; generalization; research synthesis; What Works Clearinghouse

Causal inferences about the effectiveness of programs intended to improve student learning outcomes are both important and unavoidable. Whether the evidence supporting such inferences is in some sense scientifically valid is an issue that, although hardly novel, has become a lightning rod for debate in the educational research community (Eisenhart & Towne, 2003; Howe, 2004; Raudenbush, 2005; Shavelson, Phillips, Towne, & Feuer, 2003). There appear to be two primary reasons for this. A first reason is that there is some disagreement about the nature of the evidence needed to support the causal inference that program X leads to outcome Y. The evidence necessary to answer the question, *what* is the magnitude of the effect of a program on student outcomes? is best provided by a randomized controlled experiment, the clear gold standard (although a strong quasi-experimental design may come close). However, for the evidence necessary to answer the question, *how* does a program produce an effect on student outcomes? there is no clear gold standard for a methodological approach. Indeed, a good case can be made that such questions cannot only be addressed convincingly without taking a more qualitative investigative approach (c.f. Eisenhart, 2005; Maxwell, 2004). Of the two types of causal questions delineated above, the first focuses on causal estimation, the second on causal explanation. Some would argue that a valid causal inference can be established solely by estimating a plausible causal effect; others would argue that the validity of a causal inference also requires an understanding of causal mechanism.

To the extent that we acknowledge the value in addressing both questions of causal estimation and explanation in making a causal inference, we become more attuned to a broader conceptualization

of what it means for a causal inference to have some degree of scientific validity. That is, in attempting to understand at a finer level how a cause produces an effect, we become all the more cognizant that there are mediating and intervening factors involved, and there will be occasions when the interactions with these factors make any single number intended as an estimate of an aggregate causal effect rather difficult to interpret. We begin to ask the million-dollar question: When, where, and to what extent can we generalize this causal effect? Given the difficulty of establishing and generalizing a causal inference from any single program evaluation, one might reasonably hold out hope for what a synthesis of multiple program evaluations can contribute to this venture. It is therefore very important to carefully examine the approaches taken by organizations such as the What Works Clearinghouse (WWC) and the Best Evidence Encyclopedia (BEE) in their syntheses of program evaluations, because their task is an ambitious one: They are attempting to synthesize causal inferences.

Comparing Syntheses by the WWC and the BEE

R. E. Slavin ("What Works? Issues in Synthesizing Educational Program Evaluations," this issue of *Educational Researcher*, pp. 5–15) writes that "scientifically valid and readily interpretable syntheses" constitute a "key requirement for evidence-based policy" (p. 5). The heart of Slavin's article is a discussion of the ways in which different syntheses of program evaluations compare in their attempts to come to summary conclusions about program effectiveness. Although other organizations that conduct syntheses are introduced, two principal organizations are compared throughout—the WWC and the BEE. Each organization conducts what are essentially "best-evidence syntheses" (Slavin, 1986): Criteria are established for studies eligible to be synthesized, and those studies with the strongest designs (i.e., most likely to contribute the best evidence) are given more weight when overall judgments about program effectiveness are reached in the form of numerical ratings. However, determining the strength of a study design is not so cut and dry, so what counts as best evidence differs depending on the organization doing the synthesis.

As an example of this, Table 1 presents the criteria used by the WWC and the BEE to categorize the effectiveness of a program on the basis of existing evaluations. Although the WWC and the BEE have the same number of rating categories, the criteria for placement in each category differs, and many of these differences are discussed in some detail by Slavin in his article in this issue. However, the extent to which the WWC and the BEE tend to disagree in their ratings of the same program is never explicitly

Table 1
Comparison of Program Rating Criteria for the BEE and the WWC

Best Evidence Encyclopedia (BEE)	What Works Clearinghouse (WWC)
<p>Strong Evidence of Effectiveness At least one large randomized or randomized quasi-experimental study, or multiple smaller studies, with a median effect size of at least +0.20. A large study is defined as one in which at least 10 classes or schools, or 250 students, were assigned to treatments. Smaller studies are counted as equivalent to a large study if their collective sample sizes are at least 250 students. If randomized studies have a median effect size of at least +0.20, the total set of studies need not have a median effect size this large.</p> <p>Moderate Evidence of Effectiveness One large matched study or multiple smaller studies with a collective sample size of 250 students, with a median effect size of at least +0.20.</p> <p>Limited Evidence of Effectiveness At least one qualifying study with a significant positive effect and/or median effect size of +0.10 or more.</p> <p>Insufficient Evidence Studies show no significant differences.</p> <p>No Qualifying Studies</p>	<p>Positive Effects Strong evidence of a positive effect with no overriding contrary evidence. Two or more studies showing statistically significant positive effects, at least one of which met WWC evidence standards for a strong design. No studies showing statistically significant or substantively important negative effects.</p> <p>Potentially Positive Effects Evidence of a positive effect with no overriding contrary evidence. At least one study showing a statistically significant or substantively important positive effect. No studies showing statistically significant or substantively important negative effects and fewer or the same number of studies showing indeterminate effects than showing statistically significant or substantively important positive effects.</p> <p>Mixed Effects Evidence of inconsistent effects as demonstrated through either of the following: at least one study showing a statistically or substantively important positive effect and at least one study showing a statistically significant or substantively important negative effect, but no more such studies than the number showing a statistically significant or substantively important positive effect, or at least one study showing a statistically significant or substantively important effect and more studies showing an indeterminate effect than showing a statistically significant or substantively important effect.</p> <p>No Discernible Effects No affirmative evidence of effects. None of the studies show a statistically significant or substantively important effect, either positive or negative.</p> <p>No Qualifying Studies</p>

Note. The WWC rating scheme also contains the categories Potentially Negative Effects and Negative Effects, which are parallel to the categories and criteria for Potentially Positive Effects and Positive Effects. It is unclear how such evidence would be categorized under the BEE rating scheme.

quantified in Slavin’s discussion. Table 2 remedies this to a limited extent by providing a cross-tabulation for 11 common syntheses of curricular programs in elementary and middle school mathematics conducted by both the WWC and the BEE. The correlation between the two sets of ratings is .57. In only 2 out of 11 cases was a program given the same rating. In fact, in 4 out of 11 cases, the ratings differ by more than two categories. Although in theory the ratings provided by the WWC and the BEE are both measures of a given program’s effectiveness, the results summarized in Table 2 suggest that they are measuring different things.

Slavin points to four sources of disagreement in the ways that program evaluations are synthesized in the WWC relative to the BEE:

1. Inclusion of biased outcome measures,
2. Inclusion of studies with brief program duration (less than 12 weeks),

3. Failure to weight studies by sample size, and
4. Overvaluing the contributions of randomized experiments relative to strong quasi-experiments or “randomized quasi-experiments.”

Although I generally agree with Slavin on these points, I found it curious that he does not connect his relatively muted concerns with two stronger critiques that have recently been levied at the WWC (Confrey, 2007; Schoenfeld, 2006).

In these very pages, Alan Schoenfeld (2006), who had served as a senior content advisor for the WWC’s middle school math review, described problems with the alignment of tests measuring mathematical proficiency and the curricula these tests are used to evaluate. In an article published in *Educational Evaluation and Policy Analysis*, Jere Confrey (2007), who had served as the chair of the National Research Council’s Committee for the Review of the Evaluation

Table 2
Comparison of Program Effectiveness Ratings: Mathematics Curricula

Best Evidence Encyclopedia (BEE)	What Works Clearinghouse (WWC)				
	No Qualifying Studies	No Discernible Effects	Mixed Effects	Potentially Positive Effects	Positive Effects
No qualifying studies	0	1	1	0	0
No discernible effects	0	2	0	1	0
Mixed effects	0	1	0	3	2
Potentially positive effects	0	0	0	0	0
Positive effects	0	0	0	0	0

Note. For the sake of interpretability, the category labels for the WWC have been applied to the BEE categories.

Data on the Effectiveness of NSF-Supported and Commercially-Generated Mathematics Curriculum Materials, raised similar concerns and also discussed major philosophical differences between the approach taken by the WWC to conduct syntheses of mathematics curricula and the approach taken by the National Research Council (NRC).

The purpose of the present article is to expand on two of the key arguments raised by Schoenfeld (2006) and Confrey (2007) in their critiques of syntheses produced by the WWC. Both arguments raise questions about the validity of synthesized causal inferences when little attention is given to issues of generalizability. The first argument is that the WWC overlooks the importance of construct validity in the way that both programs and test outcomes are operationalized from study to study. The second argument is that the WWC overemphasizes the internal and statistical conclusion validity of the causal effects estimated from study to study at the expense of reaching conclusions about external validity. I think that these two arguments can be placed within a broader validity context that captures Slavin's four points of disagreement between the WWC and the BEE and also point to weaknesses that are present in both synthesis approaches. I conclude with some suggestions for strengthening the validity of causal inferences in future ratings-based syntheses.

Threats to the Validity of Causal Inferences

It has been 45 years since Campbell and Stanley (1963) introduced a terminology for discussing "factors jeopardizing" the validity of causal inferences stemming from both experimental and quasi-experimental designs.

Internal validity is the basic minimum without which any experiment is uninterpretable: Did in fact the experimental treatments make a difference in this specific experimental instance? *External validity* asks the question of generalizability: To what populations, settings, treatment variables, and measurement variables can this effect be generalized? Both types of criteria are obviously important, even though they are frequently at odds in that features increasing one way jeopardize the other. While internal validity is the sine qua non, and while the question of external validity, like the question of inductive inference, is never completely answerable, the selection of designs strong in both types of validity is obviously our ideal. (p. 5)

Cronbach, writing from the perspective of a program evaluator, famously objected to Campbell and Stanley's characterization of

internal validity as being more important than external validity. Cronbach (1982) argued that for the results of a program evaluation to be relevant to a broader educational context, it becomes necessary to extrapolate a causal inference beyond the specific students, school settings, program implementations, and measurement outcomes internal to the study. If such an extrapolation is not warranted, then it would make little difference whether an observed causal effect was internally valid, as Campbell and Stanley had defined the term, because the evaluation would lack relevance for subsequent decision makers. Cook and Campbell (1979) and, more recently, Shadish, Cook, and Campbell (2002) attempted to reconcile the original Campbell and Stanley terminology and validity conceptualizations with Cronbach's objections. Their solution was to distinguish between four categories of validity threats, and these are the definitions I invoke when I write about internal and external validity in what follows:

1. *Statistical conclusion validity.* The validity of inferences about the correlation (covariation) between treatment and outcome.
2. *Internal validity.* The validity of inferences about whether observed covariation between A (the presumed treatment) and B (the presumed outcome) reflects a causal relationship from A to B as those variables were manipulated or measured.
3. *Construct validity.* The validity of inferences about the higher order constructs that represent sampling particulars across persons, treatments, outcomes, and settings.
4. *External validity.* The validity of inferences about whether the cause-effect relationship holds over variation in persons, settings, treatment variables, and measurement variables.

Although statistical conclusion validity and internal validity jointly make it possible to conclude that a causal effect has been established in a given study, construct validity and external validity are necessary before it is possible to generalize a causal effect. It is worth noting that the definition of construct validity above is consistent with the way Campbell and Stanley originally conceived of external validity, as a form of generalization, whereas the definition of external validity above captures Cronbach's view that the crux of all evaluation results is the ability to make prospective statements about the effectiveness of a program in novel contexts.

The approaches taken by both the WWC and the BEE in synthesizing evaluations of program effectiveness focus almost exclusively on reaching conclusions about the internal and statistical conclusion validity of a collection of causal inferences, not the generalizability of those causal inferences. In my view, it is the lack of careful attention being given to issues of construct validity and external validity that has led to the entirely legitimate critiques and disagreements voiced by Slavin, Schoenfeld, and Confrey about the way the WWC goes about synthesizing the findings from educational programs into numerical ratings of effectiveness.

Construct Validity

Construct validity applies to each of three fundamental elements of any educational program evaluation: the units of both assignment and analysis (usually students, teachers, or schools), the experimental treatments, and the test outcomes. The question in each case is whether the specific elements observed in an evaluation are representative of the elements that were intended. This is relatively easy to conceptualize when students are both the units of analysis and assignment. We naturally ask, What is the target population that this sample of students represents? When students have been sampled using probability methods, this question has a clear answer; when students comprise a sample of convenience, the answer becomes more ambiguous. Establishing construct validity can be especially challenging in the context of experimental treatments and test outcomes, and in the next two sections I focus on examples of such challenges as they apply to synthesizing causal inferences.

Implementing Experimental Treatments

Ratings of a program's effectiveness may well be misleading when the experimental treatments that are implemented to estimate a causal effect depart from what was intended. Slavin gives a good example of this when he criticizes the WWC for its decision to provide ratings for the effectiveness of the Daisy Quest software in increasing phonemic awareness, because (a) the duration of the program in the underlying experimental studies was less than 5 hours and (b) students in the treatment group received additional tutoring beyond the instruction unique to the Daisy Quest software. Confrey (2007) gives a different example when she discusses the WWC decision to give the program Saxon Math a rating of positive effects on the basis of a single evaluation that qualified as meeting WWC evidence standards without reservations (Peters, 1992; WWC, 2004). The study in question, Saxon Math, a standalone middle school math curriculum, had been supplemented with "the use of calculators, a weekly 50-minute computer class, and the use of cooperative learning" (Confrey, 2007, p. 205). For both Daisy Quest and Saxon Math, the intended construct of interest is a curricular program intended to affect reading and math achievement outcomes, respectively. Yet in each case, the advertised components of the program differed substantively from what was actually implemented in empirical evaluations of the program.

A related problem is that in many studies the specific program received by the control group is unspecified (Confrey, 2007). This is critically important because it is the choice of control group that makes an estimated causal effect interpretable. The

control group provides the frame of reference for an effect. For example, Schoenfeld (2006) describes two competing types of mathematics curricula: traditional and reform. To the extent these are genuine distinctions, it will matter quite a bit if the estimated effect of a math program with a reform curriculum is based on a comparison with a control group consisting of a program with a traditional curriculum or an alternative reform curriculum. Another way to think of this is that "no effect" might be a very positive evaluative outcome if big learning gains are observed for both treatment and control groups.

Testing Outcomes

The most encompassing definition of the construct validity of testing outcomes was provided by Messick (1989) when he wrote that "validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (p. 13). It is by now well understood that the construct validity of a test is something that is never established unequivocally. Theorists such as Cronbach (1988), Messick, and Shepard (1993) have instead emphasized the notion that test validation should be viewed as an ongoing process of scientific inquiry. In the latest edition of *Educational Measurement*, Kane (2006) expanded on what he has called "an argument-based approach to validity" (Kane, 1992). Kane's thesis is that test validity is a matter of degree and depends on the clarity, coherence, and plausibility of any interpretive argument that links test scores to the decisions and inferences for which they are to be used. Gorin (2006) suggests that in every test one should distinguish between the measurement construct that was intended and the construct that was actually enacted. These sorts of understandings appear to be notably absent from the synthesis efforts undertaken by both the WWC and the BEE. For the most part, the claims made by a test publisher about what a standardized test is measuring are accepted at face value, and construct validity is viewed as a static property of the test, independent of its use.

Once again, inattention to issues of construct validity can lead to misleading conclusions about a program's effectiveness. Schoenfeld (2006) discusses the results from a study by Ridgway et al. (2000) in which the mathematics performance of California students was compared on two different outcomes: the STAR test and the Balanced Assessment test. Ostensibly, both tests were measures of the general construct of mathematical proficiency. Yet upon closer inspection it is clear that the STAR test was designed to measure a considerably narrower domain of mathematical proficiency than the Balanced Assessment test. According to Schoenfeld,

The data indicated that between 70% and 75% of the students at each grade level scored equivalently (either proficient or not proficient) on both tests. However, fewer than 5% of the students scored proficient on the standards-based test [Balanced Assessment] and not proficient on the skills-oriented test [STAR], while about 22% of the students were deemed proficient on the skills-oriented test but not proficient on the standards-based test. The latter group of students, nearly a quarter of the student population, was deemed "proficient" by the state of California on the basis of the STAR test, but that ostensible proficiency may well have been an artifact of the narrowness of the test. (p. 17)

The implications of Schoenfeld's example should be clear. Imagine that we wish to evaluate the effectiveness of a mathematics program based on a reform curriculum against a control condition based on a traditional curriculum. Using something like the STAR test as an outcome measure would be likely to bias findings against the reform curriculum. Slavin cites the evaluation of Saxon Math as an example in which the bias would go in the opposite direction because the outcome measure is designed to align with the contextual classroom experiences of students exposed to the experimental treatment but not with those of the students in the experimental control. Both of these examples have an established terminology in test validity theory: Schoenfeld's example is one of *construct underrepresentation*; Slavin's example is one *construct irrelevant variance*.

The construct validity of outcome measures across program evaluations makes a big difference in the way that subsequent syntheses of these studies should be interpreted by educational stakeholders. When a school or school districts makes the decision to implement a program, they may hope in a general sense that their students learn more as a result, but under the current auspices of educational accountability systems, the bottom line will be the effect of the program on the state standardized test. For example, after consulting the evidence of what works among middle school math programs at the WWC, a school district in my home state of Colorado might decide to implement program A over program B. The expectation would be that this decision would have an effect of student achievement on Colorado's standardized tests in math. However, if the primary basis for the WWC rating of positive effects for program A comes from a study in which the test outcome contained construct irrelevant variance that favored program A, whereas the primary basis for a WWC rating of no effect for program B came from a study in which the construct of outcome measure was underrepresented, the school district will be basing its decision on faulty evidence. What any school district would want to know is whether the tests used to evaluate a given program are comparable to the tests being used to hold schools accountable. Such insights are missing from the WWC and the BEE syntheses.

Although this is not the terminology he invokes, construct validity is at the heart of Schoenfeld's (2006) critique of the WWC. In a response to Schoenfeld (Herman, Boruch, Powell, Fleischman, & Maynard, 2006), representatives of the WWC appear to claim that there are no standards available to objectively evaluate the validity of the test outcomes used in program evaluations.

For the WWC to try to adjudicate an issue about which there is no consensus in the field would be presumptuous and would immerse the WWC in a debate about "acceptable" outcome measures that the WWC could not hope to resolve. Furthermore, this task would detract from the time and resources to conduct reviews, our principal task. (p. 22)

The sentiments expressed in these two sentences are revealing. After all, there is a consensus in the field about forms of evidence that support the validity of test interpretations and uses. This consensus is embodied by chapter 1 of the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on

Measurement in Education, 1999; hereafter referred to as the Test Standards for Validity). If it is presumptuous to use the framework of the Test Standards for Validity to rate the quality of the outcome measure used in a program evaluation, then it is just as presumptuous to use the WWC Standards framework to rate the quality of an evaluation study design. That Herman et al. view this activity as "detracting from the time and resources to conduct reviews" suggests that they place little or no value on the importance of generalizing causal inferences.

External Validity

When construct validity has been established in a given study, it becomes possible to draw conclusions along the lines of

Program X had the effect Z on outcome Y. On the average this effect applies to (a) students/teachers/schools/communities with the following characteristics . . . , (b) a program with the following implementation characteristics . . . , and (c) a test with the following measurement characteristics. . . .

In attempting to further generalize this inference, we ask, Can we expect the program to have the same effect on student outcomes if we vary

- a. the types of students, teachers, schools, and communities involved?
- b. the way the program is implemented? and
- c. the outcomes to be measured?

It can be difficult to address each of these questions on the basis of evidence from a single evaluation study. In such cases, generalization follows from *interpolations* of estimated effects for different student subgroups, program implementations, and outcome measures found in a given study.

In many studies with quasi-experimental designs, researchers have no control over assignment to treatment conditions, but they have considerable control over the sampling of the units of analysis. Some of the best available data for such studies can be found in the large-scale surveys sponsored by the National Center for Education Statistics. Although randomized experiments are impossible with these surveys, the trade-off is that the available samples of students and schools are both large and nationally representative. In studies such as these, researchers have the statistical power both to estimate an aggregate causal effect and to explore a variety of important interactions. For example, I used data from the National Educational Longitudinal Study (NELS) of 1988 to estimate the effect of commercial coaching programs on the SAT and ACT college admissions exams (Briggs, 2001, 2004). In doing so, I controlled for potential confounders, such as prior academic achievement, demographic characteristics, and proxies for student motivation, using a linear regression model. But, in addition, because the available sample included more than 3,000 students, I was also able to compare these aggregate effect estimates with those for student subgroups defined as a function of different demographic characteristics. This led to one very striking finding: Although the aggregate effect of coaching on the verbal section of the SAT was about 10 points, the effect specific to Hispanic students was -18 points. The effect of coaching did not generalize among the racial/ethnic subgroups in the NELS sample.

This is not to suggest that the ability to interpolate causal effects is unique to quasi-experiments—indeed, one of the key findings from the Tennessee STAR experiment was that the effects of class-size reduction were largest for minority students in urban schools (Finn & Achilles, 1999). But randomized experiments on the scale of the STAR experiment are the exception, not the rule, in educational research. In general, randomized experiments tend to be smaller and more homogeneous; quasi-experiments tend to be bigger and more heterogeneous.¹ Because of this, on balance, a large quasi-experiment will deliver more bang for the buck than a small randomized experiment when the name of the game is causal generalization. However, when the elements of a study are relatively homogeneous, generalization follows from extrapolations beyond the particulars of a study's internal context. The extrapolations from any one study will usually rely on logical argument rather than on statistical inference. This points to a clear advantage of synthesizing causal inferences across multiple studies: the ability to test the extrapolations made on the basis of individual studies empirically.

Given the objective of both estimating and generalizing an estimated causal effect, the apparent ideal would be to have a collection of randomized experiments with considerable heterogeneity among the students, treatments, and outcomes across studies. However, when there are very few of these studies available or when the studies focus on a very restricted domain of experimental conditions, having a large collection of quasi-experiments becomes all the more important. This is because, in my view, the central purpose of a synthesis differs from the central purpose of a program evaluation when the context is causal inference: The goal of the one-shot evaluation is to estimate a valid causal effect; the goal of a synthesis should be to determine if that causal effect is generalizable.

Slavin alludes to empirical evidence that in evaluations of math and reading programs results based on high-quality quasi-experiments tend to mirror those based on randomized experiments. Slavin also points out that the findings from a randomized experiment may still be biased when the experiment involves a small sample size. When this is coupled with the usual threat of differential attrition across experimental groups, I am reminded of one of Paul Holland's (2004) more colorful aphorisms: "A randomized controlled experiment is just a quasi-experiment waiting to happen." Slavin argues that all things being equal, randomized experiments remain preferable as a safeguard against selection bias when estimating a causal effect.² This is certainly true, but echoing Cronbach (1982), it is not clear to me that an unbiased causal effect has much relevance if it cannot be generalized. And if a primary purpose of conducting a synthesis is to generalize, then perhaps the WWC and the BEE should differentiate two sets of evidence standards: one for the collection of studies used to establish an unbiased causal inference and another for the collection used to generalize that inference.

Discussion

According to Slavin, good syntheses of program evaluations are critical because "educational policy cannot support the adoption of proven programs if there is no agreement on what they are" (p. 5). Although I share Slavin's sentiment that educational policy should be based on empirical evidence, I take issue with the

implicit notion that programs should only be classified with respect to evidence that they have been, in some limited sense, "proven to work." It would seem that, philosophically, Slavin is taking the same basic position as the WWC: The focus of syntheses must be on what has worked, that is, programs for which there is evidence of an aggregate effect that is internally valid. I would argue that such evidence, although certainly important, is necessary but not sufficient for those stakeholders enacting educational policies. What the superintendent of a school district wants to know is not so much what *has* worked but what *will* work. To be relevant, a good synthesis should give policy makers explicit guidance about program effectiveness that can be tailored to specific educational contexts: When and where will a given program work? For whom will it work? Under what conditions will it work the best? For causal inferences to be truly valid, both causal estimation and generalization should at the very least be given equal weight. This is entirely consistent with the view of scientific research in education as endorsed by the NRC (2002) in the book *Scientific Research in Education*:

Scientific Principle 5: Replicate and Generalize Across Studies

Scientific inquiry emphasizes checking and validating individual findings and results. *Since all studies rely on a limited set of observations, a key question is how individual findings generalize to broader populations and settings* [italics added]. Ultimately, scientific knowledge advances when findings are reproduced in a range of times and places and when findings are integrated and synthesized. (p. 4)

It is informative to compare the philosophical approach of the WWC to that which is generally taken by the appointed committees of the NRC when they conduct syntheses of program evaluations. In the context of synthesizing the findings from evaluations of mathematics curricula, Confrey (2007) describes the difference between the two approaches as follows:

NRC takes the position that the conduct of evaluation must be informed by careful study of existing evaluations of mathematics curricula, leading to a proposed definition of "scientifically established as effective"; whereas WWC claims general and prior knowledge that quality evaluations are achieved through a particular methodology applied to the review of any form of regular educational intervention. (p. 197)

On the basis of his previous writings (Slavin, 1986, 2002, 2004), I would guess that Slavin's views fall somewhere in between these two philosophies. A strong feature of the syntheses conducted by Slavin's BEE is that for any given topic area (i.e., middle school math, comprehensive school reform) a more comprehensive report is provided below the link to an "educator's summary" document that includes program ratings. In many cases, these reports are subsequently published in peer-review journals. A perusal of these reports indicates some attention being given to issues of causal generalization (although not in a very explicit manner), certainly more attention than is given in the reports released by the WWC.

However, there can be little question that in both the syntheses conducted by the WWC and the BEE, the prospective evaluation of causal generalization takes a distant backseat to the retrospective evaluation of estimated causal effects. It seems to me that this is an inefficient use of time and energy. For example, in

synthesizing evaluations of the program Success for All, the WWC (2007) gathered a total of 74 studies that investigated the effect of this program on student achievement. Out of these 74, only 7 were suitable for inclusion according to the WWC evidence standards. Is there really nothing to be learned from the 67 studies that were excluded from the formal synthesis? In particular, one might expect many of these nonexperimental studies to provide insights about the issues of construct validity that I have raised above, issues that are, unfortunately, often given superficial treatment in experimental studies. If the trouble was taken to collect (and hopefully) read the full set of studies in the first place, surely it is possible to cull some information that would provide both a broader context and deeper insights for the subset of studies being synthesized quantitatively.

I do not mean to give the impression that the sort of synthesis efforts undertaken by the WWC (and for that matter, the BEE) are hopelessly flawed. Even in their current form, the WWC syntheses are valuable if only because they do indeed serve as a well-publicized clearinghouse for educational stakeholders to consult before they implement an educational program. The methods applied by the WWC in its reviews, whether one agrees with them or not, are relatively transparent. Most important, the WWC makes it a point to note that its reviews are in a constant state of flux, subject to revision as evidence in the form of new evaluative studies becomes available. What makes me uncomfortable is the idea that any single number should be used to indicate the extent to which an individual program *works*, using a conception of the term that is limited and retrospective. One way to mitigate this concern is to employ more than one set of ratings. In the same way that one would advocate multiple measures that draw conclusions about the quality of an individual student, if numbers are to be attached to individual programs, they should also be based on multiple measures.

I conclude then, with six specific suggestions for improving the practice of synthesizing causal inferences across program evaluations when numerical ratings are the desired result:

1. Explicit distinctions should be made between (a) establishing the magnitude and direction of the effects attributed to a given educational program and (b) the generalizability of the estimated effects across different combinations of students, settings, treatment implementations, and outcome measures.
2. If ratings are to be given about the effectiveness of a given program, a parallel set of ratings should be given about the generalizability of these findings.
3. Ratings of generalizability should have two components, one that focuses on construct validity and another that focuses on external validity.
4. The appropriate framework to be used to evaluate the construct validity of a given testing outcome comes from the Test Standards for Validity, which divides validity-related evidence into five categories:
 - a. Evidence based on test content,
 - b. Evidence based on response processes,
 - c. Evidence based on internal structure,
 - d. Evidence based on relations to the variables, and
 - e. Evidence based on test consequences.

5. In establishing a rating for the external validity of a causal inference, all relevant student, treatment, and outcome categories should be examined for evidence of consistency or inconsistency in estimated causal effects.
6. All program syntheses should conclude with a section detailing gaps in our understanding about both the effectiveness of the program and the generalization of its effectiveness.

NOTES

¹It is worth noting that both the What Works Clearinghouse in its syntheses and Slavin in his article in this issue appear to conflate sample size in and of itself with external validity.

²Slavin also makes the peculiar argument that “selection bias may balance out in the long run, over many studies” (p. 8). There is no statistical reason to expect this to be so—selection bias is by definition a random event. If it is positive in one study, it is more likely to be positive in others as well.

REFERENCES

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Validity*. In *Standards for educational and psychological testing* (pp. 9–24). Washington, DC: American Educational Research Association.
- Briggs, D. C. (2001). The effect of admissions test preparation: Evidence from NELS-88. *Chance*, 14(1), 10–18.
- Briggs, D. C. (2004). Evaluating SAT coaching: Gains, effects and self-selection. In R. Zwick (Ed.), *Rethinking the SAT: Perspectives based on the November 2001 Conference at the University of California, Santa Barbara*. New York: RoutledgeFalmer.
- Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.
- 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.
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.
- Cronbach, L. (1982). *Designing evaluations of educational and social programs*. San Francisco: Jossey-Bass.
- Cronbach, L. (1988). Five perspectives on the validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3–15). Mahwah, NJ: Lawrence Erlbaum.
- Eisenhart, M. (2005). Hammers and saws for the improvement of education research. *Educational Theory*, 55(1), 245–261.
- Eisenhart, M., & Towne, L. (2003). Contestation and change in national policy on “scientifically based” educational research. *Educational Researcher*, 32(7), 31–38.
- Finn, J., & Achilles, C. (1999). Tennessee’s class size study: Findings, implications, misconceptions. *Educational Evaluation and Policy Analysis*, 21(2), 97–110.
- Gorin, J. (2006). Test design with cognition in mind. *Educational Measurement: Issues and Practice*, 25(4), 21–35.
- 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.
- Holland, P. W. (2004, April). *Evidence for causal inference in education research*. Paper presented at the American Educational Research Association Presidential Invited Session on Inference, Evidence and Scientific Research, San Diego, CA.
- Howe, K. (2004). A critique of experimentalism. *Qualitative Inquiry*, 10(1), 42–61.

- Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112(3), 527–535.
- Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education and Praeger.
- Maxwell, J. A. (2004). Causal explanation, qualitative research, and scientific inquiry in education. *Educational Researcher*, 33(2), 3–11.
- Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education and MacMillan.
- National Research Council. (2002). *Scientific research in education*. Washington, DC: National Academy Press.
- Peters, K. (1992). *Skill performance comparability of two algebra programs on a eighth grade population*. Unpublished doctoral dissertation, University of Nebraska, Lincoln.
- Raudenbush, S. W. (2005). Learning from attempts to improve schooling: The contribution of methodological diversity. *Educational Researcher*, 34(5), 25–31.
- Ridgway, J., Crust, R., Burkhardt, H., Wilcox, S., Fisher, L., & Foster, D. (2000). *MARS report on the 2000 tests*. Palo Alto, CA: Silicon Valley Mathematics Assessment Collaborative.
- 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.
- Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.
- Shavelson, R. J., Phillips, D. C., Towne, L., & Feuer, M. J. (2003). On the science of education design studies. *Educational Researcher*, 32(1), 25–28.
- Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education*, 19, 405–450.
- Slavin, R. E. (1986) Best-evidence synthesis: An alternative to meta-analytic and traditional reviews. *Educational Researcher*, 15(2), 5–11.
- Slavin, R. E. (2002) Evidence-based education policies: Transforming educational practice and research. *Educational Researcher*, 31(7), 15–21.
- Slavin, R. E. (2004) Education research can and must address “what works” questions. *Educational Researcher*, 33(1), 27–28.
- Slavin, R. E. (2008). What works? Issues in synthesizing educational program evaluations. *Educational Researcher*, 37, 5–14.
- What Works Clearinghouse. (2004, October 28). *Detailed study report on Peters (1992)*. Retrieved April 25, 2005, from http://www.what-works.ed.gov/PDF/Peters_1992_Detailed_Study_Report.pdf
- What Works Clearinghouse. (2007, August 13). *Beginning Reading Intervention report: Success for All*. Retrieved December 15, 2007, from http://ies.ed.gov/ncee/wwc/pdf/WWC_Success_All_BR_081307.pdf

AUTHOR

DEREK C. BRIGGS is an assistant professor at the University of Colorado, School of Education, 249 UCB, Boulder, CO 80309; Derek.Briggs@colorado.edu. His research focuses on building sound methodological approaches for the valid measurement and evaluation of growth in student achievement.

Manuscript received December 23, 2007

Accepted January 4, 2008