



Comments on Slavin

# Through the Looking Glass: Experiments, Quasi-Experiments, and the Medical Model

Finbarr Sloane

Slavin (2008) has called for changing the criteria used for the inclusion of basic research in national research synthesis clearinghouses. The author of this article examines a number of the assumptions made by Slavin, provides critique with alternatives, and asks what it means to fully implement the medical model in educational settings.

**Keywords:** construct validity; critique; external validity; fixed and random effects; group-randomized experiments; quasi-experiments; randomized experiments

In his article “What Works? Issues in Synthesizing Educational Program Evaluations” (this issue of *Educational Researcher*, pp. 5–14), Robert E. Slavin calls for a set of uniform criteria to be used across synthesis clearinghouses for the inclusion of evaluative studies of instructional programs. He highlights differences in criteria used by a number of these clearinghouses including the What Works Clearinghouse, the Comprehensive School Reform Quality Center, the Best Evidence Encyclopedia, the Evidence for Policy and Practice Information and Co-ordinating Centre in Great Britain, and the Campbell Collaboration. Furthermore, he elaborates a promising set of common criteria, arguing coherently for the inclusion of each. The suggested criteria include the following:

- a. Randomized designs should be preferred to matched designs, but large, well-controlled matched designs contribute important information.
- b. 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.
- c. Among matched studies, prospective studies should be strongly preferred to retrospective comparisons. If there is a sufficient number of better quality studies, retrospective studies should be excluded.
- d. Small studies can have highly variable effects and suffer more from publication bias than their larger counterparts. They often have confounds with school, teacher, and class effects. Larger studies should be preferred. Weighting by sample size may be used.

- e. Matched studies in which pretests are not given, and those in which pretest differences are less than 50% of a standard deviation, should be excluded. Randomized experiments without pretests are acceptable if attrition is low and equal between experimental and control groups.
- f. Studies of less than 12 weeks in duration should be excluded.
- g. Measures inherent to or potentially biased toward experimental treatments should be excluded.
- h. Program ratings should be created according to the strength of evidence, balancing median effect size, number of studies, and quality of research design. The outcomes of large, randomized experiments should be strongly emphasized.

The goal of this article, simply put, is to evaluate the arguments with which Slavin underscores his suggested criteria. In many cases, I agree with him; in others, I do not. In all cases, I will elaborate on the argument he makes for criterion inclusion by outlining some of the hidden assumptions being made in favor of the specific criterion.

One could say that Slavin’s call for changes in some of the quality characteristics for study inclusion in the aforementioned clearinghouses, through his negotiated model that he proposes should be common across the synthesis houses, drew somewhat on his being unhappy that some of his own work has been perceived as not being of the quality necessary for inclusion in the highest category of the What Works Clearinghouse. In arguing his case, Slavin inevitably climbs a slippery intellectual slope around these issues. Of course, it would be easy, and here much too easy, I believe, to see this call for change as little more than self-serving. However, this would be unfair and also would undermine the critical intellectual questions that he raises. Suffice to say that Slavin engages his “climb” with due care and attention to detail in the provocative style that has become a hallmark of his recent policy and methodology statements (Slavin, 2002). As such, it is difficult to disagree with many of his insightful comments.

To begin, there is a peculiar dynamic inherent in the article. Although Slavin argues for a new and common set of criteria for the inclusion of intervention studies, in so doing he also focuses attention on quality criteria, not just for inclusion but also for the conduct of summative research on educational innovations. Slavin notes that this call for increased quality in educational research is not new, and he is correct. He is also correct in suggesting that the urgency associated with the call has been heightened in recent times

(Education Sciences Reform Act, 2002). Over time, the same call for quality research in education has been made by many researchers, perhaps most eloquently by Bloom (1972) when he wrote:

In education, we continue to be seduced by the equivalent of snake-oil remedies, fake cancer cures, perpetual-motion contraptions, and old wives' tales. Myth and reality are not clearly differentiated, and we frequently prefer the former to the latter. . . . We have been innocents in education because we have not put our own house in order. We need to be much clearer about what we do and do not know so that we don't continually confuse the two. If I could have one wish for education, it would be the systematic ordering of our basic knowledge in such a way that what is known and true can be acted on, while what is superstition, fad, and myth can be recognized as such and used when there is nothing else to support us in our frustration and despair. (p. 332)

In my view, Slavin, either explicitly or implicitly, makes a number of basic assumptions that warrant further deliberation. These include but are not limited to the following:

1. That the cost of quality research, especially at scale, is perhaps prohibitively expensive. Consequently, we need to generate a commonly shared and more flexible set of quality criteria for inclusion in research syntheses so that we do not exclude critically important research.
2. That in practice, as well as in theory, studies of curricular interventions in real settings can and should allow for full random assignment (Suggestions a and b, above). By full random assignment I mean assignment settings where treatments are assigned to students at random and then treatment classrooms are assigned to teachers at random. However, Slavin argues that it is often more convenient to employ clustered random assignment. Here treatments are assigned at random to classrooms (or schools). He notes that even using the simpler of these two random assignment processes, there can be considerable and prohibitive cost constraints due to the large number of clusters (classrooms or schools) required for treatments that produce reasonable effect sizes.
3. That internal validity is more important than external validity (Suggestions a, b, and e, above, and related to Assumption 4, below) and that issues of external validity are addressed through the combined effects of studies available in the specific clearinghouse.
4. That above all else we should be interested in main effects! And when this is the case, we can ignore significant between-group variance. As such, we can also ignore interactions between treatments and the individual characteristics of the participating students (embedded in Suggestion b and related to Assumption 3, above).
5. That it is reasonable to omit validity issues in measurement when choosing studies for inclusion as long as the researcher uses a "known" standardized instrument (Suggestion g).
6. That the duration of the treatment is a critical design factor (Suggestion f).

This list of assumptions focuses attention on a few of the many undergirding assumptions Slavin makes in his argument. I will

attend to each of them, although not separately in every case due to space limitations.

### **What Is Lost When We Omit the Nesting of Data and Ignore the Possibility for Random Effects?**

It is commonly held that teachers, even when scripted, implement instructional treatments with some degree of variability. In real life, treatments change in their adoption and adaptation by teachers. This may also be true in the experimental setting especially when a large number of classrooms is sampled. Generally speaking, it is standard practice in psychology to assume that treatments have fixed effects. This results in studies having large power to detect very small effects, including some that are of little practical consequence. Ignoring the nesting of educational data and the possibility for random effects is debilitating, as Cronbach (1976) pointed out more than 30 years ago: "The majority of studies of educational effects . . . have collected and analyzed data in ways that conceal more than they reveal. The established methods have generated false conclusions in many studies" (p. 1). More than that, Slavin's assumption puts at odds the traditions of psychology, in which treatments are fixed effects, and real life, in which treatments vary by teachers, students, schools, school districts, and so on. The issue of random effects does not raise its ugly head when one tests the effect of a treatment with only two classrooms, a control and treatment. However, this assumption needs to be tested with studies in school settings.

Let me review: Psychologists treat all categorical, explanatory variables (e.g., competing treatment, or treatment versus control groups) as if they were the same. This was certainly what R. A. Fisher had in mind when he invented the analysis of variance in the 1930s (Fisher, 1937). Eisenhart (1947) realized that there were actually two fundamentally different sorts of categorical, explanatory variables; he called these fixed and random effects:

- Fixed effects influence only the mean of  $y$ .
- Random effects influence only the variance of  $y$ .

A random effect should be thought of as coming from a population of effects: The existence of this population is an extra assumption. Statisticians speak of prediction, rather than estimation, of random effects. In contrast to random effects, from which we want to make predictions about the population of random effects, fixed effects are unknown constants to be estimated from data. Random effects govern the variance-covariance structure of the response variable. The key point is that observations affected by random effects are not independent (e.g., students nested within classrooms). Observations that contain the same random effect are correlated, and this contravenes a fundamental assumption of standard statistical models: the independence of errors.

Many kinds of statistical work involve studies conducted at a range of spatial scales. For example, in education we might investigate the effects of curriculum on people as individual students, instructional groups, classrooms, or schools. Naturally, the questions we ask at different scales would concern different processes. What is important to understand is that variability accumulates with the broadening of spatial scale; thus, variation between instructional groups and classrooms contributes to the variation

in the school as a whole, and this is what Slavin in his article in this issue asks us to omit. Citing Raudenbush and Bryk (2002), he argues that clustered, random assignment with analysis at the level of the student provides unbiased effects sizes at the level of the student. However, he fails to note that the confidence intervals would be inaccurate. One might ask, what is the real value of the mean effects? What do we lose when we are unable to accurately estimate the appropriate confidence intervals? Moreover, his emphasis on mean effects leaves us in the dark about what works for whom and when. Surely questions of external validity are critical if we are to suggest that school districts purchase these “tested” materials for all their students.

Treatment effects as fixed or random accordingly constitute two separate models. The two contrasting models can be described here: In Model I, differences between means are ascribed entirely to a fixed treatment effect, so any data point can be decomposed as follows:

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}.$$

It is the overall mean,  $\mu$ , plus a treatment effect,  $\alpha_i$ , plus deviation,  $\epsilon_{ij}$ , that is unique to that individual and that is drawn from a normal distribution with mean of zero and variance  $\sigma^2$ . The variance is assumed to be constant across all levels,  $j$ , which are assumed to differ only in their means. Because variance is constant, there is only one error term, based on  $\sigma^2$ .

Model II is subtly different. Here the model is written as

$$Y_{ij} = \mu + U_i + \epsilon_{ij}.$$

This model looks the same as Model I but for  $U_i$ . The distinction is critical. The  $\alpha_i$  and  $U_i$  are fundamentally different. The  $\alpha_i$  refer to fixed levels of treatment, whereas  $U_i$  represent a sample from a population of variances. The  $\alpha_i$  assume that when an educational effect is implemented with each of 20 teachers say, that each teacher enacts that curriculum in his or her classroom in exactly the same way and generates the same effects (irrespective of student intake as long as randomization has occurred). This is a very large assumption for educational researchers to make.<sup>1</sup> In Model I, the statistician’s burden is to randomly assign individuals to treatment groups or to a priori selected, already generated categories, that is, fixed and clearly distinctive categories. The main interest is in the value of the differences between the means of the various factor levels. In Model II, the factor levels are different from one another, but the experimenter did not make them different. They were selected from a much larger pool of factor levels that exhibit variation beyond the control of the experimenter (e.g., when teachers do not cover all of the curricular material, when they cover the material with different emphases, etc.). We call it random variation and call the effects random effects. The focus is not differences between means of factor levels but rather the way that factors influence the variance of the response.

Because the random effects,  $U_i$ , come from a larger population of possible random effects, there is not much point in concentrating on estimating the means of our small subset,  $k$ , of factor levels and no point at all in comparing individual pairs of means for factor levels. It is theoretically and empirically better to recognize them for what they are—random samples from a much

larger population—and to concentrate our intellectual efforts on their variance,  $\sigma^2_u$ . This added variation is caused by differences between levels of the random effects. Then, ANOVA becomes about estimating the size of this variance and working out its proportional contribution to the overall variation (and, with the help of multilevel modeling tools, to then account for or model this variation) rather than the significance of simple mean differences.

Statistically, the central issues involved in Model II fall into two broad categories: (a) questions about experimental design and the management of experimental error (e.g., Where does most of the variation occur, and where would increased replication be most profitable?) and (b) questions about hierarchical structure and the relative magnitude of variation at different levels within the hierarchy—studies on nested spatial scales (see Hedges & Hedberg, 2007). Conceptually, evaluative issues in education often require the combination of Model I and Model II. As educational researchers, we are interested not only in estimating the fixed effects but also in modeling the variation that occurs in Model II. That is, we are interested in the combination of these models through the application of mixed model ANOVA. In so doing, we move the conversation from simple and perhaps simplistic notions of main effects to understanding how these main effects manifest themselves in real classrooms. We change the basic research question from what works to what works for whom and in what contexts (questions that Slavin appears willing to ignore at this time). Slavin offers us little conceptually other than to analyze data at the level of the student. Alternatively, to reduce cost and improve quality, as recent articles have shown, we can reduce the number of groups in clustered, randomized experiments through the use of blocking as part of the research design (Bloom, Richburg-Hayes, & Black, 2007; Raudenbush, Martinez, & Spybook, 2007).

### Validity of Measurement Inferences

When we examine the effectiveness of curricula on student outcomes, measures are of critical importance. Slavin found that a measure based on the curriculum to be studied is more sensitive to curricular effects than are standardized measures. This is not new. Walker and Schaffarzick (1974) in a reanalysis of data from several curriculum evaluation efforts carefully examined outcome measures used in a series of studies. Specifically, they identified studies in which the outcome measure was aligned with the curriculum and those that were not. Unsurprisingly (but perhaps reassuringly), they found that students did better when tested on the content they had studied and less so when not. They concluded, “What these studies show, apparently, is not that new curricula are uniformly superior to old ones, though this may be true, but rather that different curricula are associated with different patterns of achievement” (p. 97).

In contrast, Slavin argues that when we evaluate curriculum effects we should always use standardized tests. He opposes the use of tests that show bias toward curricula. Also, he notes that a particular study of mathematics displayed larger effect sizes by one researcher when he used his own curricularly sensitive test compared to other studies of the same curriculum using standardized tests. Furthermore, he reports that the former study was included in the What Works Clearinghouse at the highest level because it used random assignment, albeit at small scale. However, in my opinion,

Slavin confuses the issue of inclusion (due to randomization) with the issue of validity.

To my mind, using more than one measure would improve almost any evaluative effort. It is critical to know what the curriculum is and is not actually delivering! Knowing that a curriculum meets its intended intellectual goals without damaging the possibility of reaching other goals is likely to be more important than knowing one answer alone. Invoking a “do no harm” attitude is sensible. But do not be confused: High-quality measurement is difficult and costly. The need for such measurement, particularly in mathematics and science education research, is now greater than ever before (see the calls from the Institute of Education Sciences and the National Science Foundation in this regard). This may be less true for measures of early reading skills, as the measurement and research literature align quite well. Generating tests that are curricularly sensitive and that allow for valid inferences is no small task (Brennan, 2006). The November 2007 issue of *Educational Researcher* is devoted to this concern, indicative of still unresolved and central questions about construct validity for the psychometric community. But if the “homemade” measure is of high quality—and in any given instance this may be a very big assumption—it makes no sense to throw the baby out with the bathwater. Slavin advocates that an evaluative researcher use a standardized measure over any other. Consequently, he assumes (a) that such a measure exists and (b) that the measure affords the opportunity for valid inferences to be drawn about student knowledge relative to the curriculum being evaluated. As noted above, this may be less true for at least some mathematics and science curricula than it is for reading.

### **Moving Beyond the Rhetoric of Randomized Controlled Trials**

I raise the idea of randomized controlled trials (RCTs) not to disparage but rather to illuminate some of the rhetoric around them. I draw on a presentation I made at the University of Maryland in 2005 titled “What Can Mathematics Education Learn From Medical Research?” describing the full RCT model used in parts of the medical community (Sloane, 2005; for further details, see American Statistical Association, 2007). I showed where this model aligns with current practice in educational research, where it does not, and where we have much more work to do if we are to use this particular model optimally. The focus in my response is on why the questions Slavin raises and the resolutions he offers are, in the long run, inadequate to the task he presents to us. In sum, Slavin asks us to mix results from quasi-experiments with those from clustered and full randomized trials. The question I pose here is, Is he asking for too much?

#### *The Full RCT Model Used in Drug Efficacy Research*

The complete RCT model<sup>2</sup> engages a large and well-articulated number of phases that, in general, work reasonably well for the medical community and that may not work as well in education. I briefly outline them here: Pre-Phase I studies present the plan and data collection protocol; Phase I studies establish feasibility; Phase II studies demonstrate initial efficacy<sup>3</sup> and improvement over historical norms; Phase III studies confirm efficacy with comparative randomized trials; and finally, Phase IV studies follow up in the real world of scaling. If we are to adopt this model

in education then we need to adopt the complete model. In contrast, Slavin tries to move some Phase II studies to the level of Phases III and IV.

In Pre-Phase I studies, if the clinical trial is to evaluate a new drug, the first step is an action plan called the Investigational New Drug (IND) Application that is presented to the Food and Drug Administration (FDA). This application contains everything known about the therapy, including all the data from laboratory and animal tests. If the FDA feels that the therapy might possibly benefit people, it approves the IND, and the first phase of clinical trials can begin.

Phase I studies are used to establish feasibility—in education this can be seen as roughly equivalent to early design research studies. These are nearly always constructed at a single institution. The focus is on the delivery mechanism, there is a search for interactions, and they are generally not randomized. They are studies of small sample sizes, with a dosage focus. They illuminate how much of the drug is needed to see an effect. In general, the patients who participate are local, and there is no effort to randomize, as this first phase looks specifically at dosage.

This contrasts significantly with the conduct of educational research. There is an important caveat in educational research because the researcher is trying to minimally examine two dosages simultaneously, making the inferential space more difficult. I say minimally because some dissemination models require researchers to train trainers, who ultimately train teachers, the effect of a treatment; it is filtered through several levels. We need to understand what doses should be given to the teacher and how the dosage should be delivered before the teacher can appropriately dose the student. Keep in mind that medical researchers conducting Phase I studies assume that the early innovation work has been conducted. Moreover, Phase I studies cannot be conducted without this downstream pipeline of innovation. Figuring out these dosing<sup>4</sup> issues is critical in mathematics and science education given the number of teachers in U.S. schools who are currently teaching out of field without appropriate training in content. Consequently, dosing may not work for education either as metaphor or in practice.

Phase II studies of initial efficacy are again, generally speaking, conducted at a single research institution. They are conducted with participants from the defined condition of interest in order to refine the delivery mechanism and to evaluate endpoints of interest. Results are compared to historical norms on prespecified hypotheses, and again, no randomization to treatment and control groups is considered necessary. A Phase II trial provides preliminary information about how well the new treatment works and generates more information about costs and benefits. This is the level of research that Slavin suggests is an adequate substitution for randomized trials in education.

Phase III trials are used to confirm or test efficacy.<sup>5</sup> They occur in a multi-institutional setting with standardized procedures. They are conducted with samples drawn from well-defined populations. They require large sample sizes and subset analyses. They have well-defined endpoints and require randomized comparisons. They include serious study controls (double blinding, where neither the doctor nor the patient knows who is actually getting the treatment). Finally, they are monitored with great care.

These trials compare a promising new drug, combination of drugs, or procedure to the current standard therapy. Phase III

trials typically involve large numbers of patients from doctors' offices, clinics, and cancer centers nationwide. The reason that the Phase III clinical trial has been initiated is that the superiority of one treatment over the other has not yet been firmly established. If you participate in a Phase III treatment trial, you are likely to be randomized (assigned by chance) to a group receiving either the new intervention or the standard intervention. Neither you nor your physician chooses whether you get the new intervention or the standard treatment.

Phase IV follow-up studies require real-world application and long-term follow up. They refine practices and can discover new applications. Some use the term *Phase IV* to include the continuing evaluation that takes place after FDA approval, when the drug or treatment procedure is already on the market and available for general use. This is also called a post-marketing surveillance study.

We cannot avoid doing the heavy intellectual lifting necessary for a model of this type to be successful in the enterprise that is educational research. Again, Slavin suggests that quasi-experiments give us enough information about educational interventions. However, we cannot have our cake and eat it too: Quasi-experiments are not randomized trials. Slavin acknowledges this situation when citing Heinsman and Shadish (1996). Medical researchers have done, and continue to do, their own heavy lifting. Keep in mind that they have the financial backing for this to occur. The current position in medical research allows for at least 10 studies in Phase I for every study in Phase II; the relationship is constant across phases. That means that at Phase I medical researchers conduct 1,000 studies to produce 1 Phase IV study. The model can only work if a well-funded flow of studies across each phase is conducted in educational settings that have also worked through the thorny intellectual issues of dosage. Not all ideas in medical research reach fruition; in fact, few do. This research community has shared standards for quality at each phase. In stark contrast to educational research, researchers at Phase III are considered no more scientific than those researchers conducting work at Phase I. If anything, the opposite is true. A representative of the National Institutes for Health<sup>6</sup> has put it this way in a public forum: "Oh, we would not want the scientists to conduct the trials; they would mess them up." Here the basic scientific insights occur in the earlier phases of the work. The statistical insights occur in the later phases. There is a nice wedding of at least two sciences, and they should not be confused as a single science. For example, the basic chemistry used to generate the action plan, also called the IND, looks nothing like the science of statistics. The conduct and functioning of the full RCT model described above (American Statistical Association, 2007) draws on the skills of at least two differently trained sets of scientists. Again, this is not the case in educational research.

### Summary of Points of Agreement and Disagreement

There is much in Slavin's article with which I agree. In particular, I agree with his comments regarding the sensible use of criteria for inclusion at the highest levels of the What Works Clearinghouse. For example, studies of short duration and studies conducted on small samples of students and classrooms may not deserve inclusion just because of random assignment. Small studies should be

excluded when the purpose of the Clearinghouse is external validity. Prospective studies are likely to be more accurate. When conducting quasi-experiments, pretest differences are of critical importance. Randomized experiments without pretest data are acceptable only inasmuch as the sample size is adequate given estimates of effect sizes. Although acceptable, we can gain power through blocking on such pretest data (Raudenbush et al., 2007). Studies of this type are expensive, but the cost cannot be avoided if we want to do high-quality research in this paradigm. Slavish adherence to simple rules is at best suboptimal and at times downright dangerous. Pragmatism must have some influence.

There are a number of things on which Slavin and I disagree. I would be professionally remiss not to mention his complete disregard for issues of validity in measurement as presented in his article. His other suggestions will serve our community and the policy community quite well insofar as he acknowledges that they can only serve in a stopgap sort of way. The fundamental issue is that for one reason or another we have taken on a model that does not perfectly fit our needs; we have decided, and I believe somewhat naively, what science is from the perspective of that model; we continue to call work at some phases of the model nonscientific; and we have grossly underfunded the needed research endeavors. Slavin argues that millions of dollars are at stake as a consequence of the rankings of studies in certain clearinghouses. I am sure he is correct. Surely the intellectual lives of millions of students are really at stake. The external validity of our research endeavors should not be ignored if we want to provide "success for all" students. If we want a system to produce scientific research in education, it needs to be adequately thought through in all of its phases; it needs to be funded at a level commensurate with the problem at hand; and we need to overcome the current, trivial battles about who's on first. Not only do we need a research infrastructure, but we need a parallel intellectual infrastructure to support the training of new researchers and the ongoing professional development of current researchers.

I thank Slavin for his provocative remarks in that they have spurred me, with the invitation of the editors, to write this article. As with all quantitative research, the proof of the intellectual pudding will be in the empirical eating. Let the debate begin.

### NOTES

I thank Sam Green for answering a number of questions I posed about unbiased estimators of the mean. I thank Brandon Holding and Dan Battey for their editorial input. I am currently in receipt of funding from the National Science Foundation (NSF 06116306). The comments presented here in no way represent the position of the Foundation. Finally, I thank Gene Glass for taking an interest in my research work and in encouraging me to put my thoughts on paper.

<sup>1</sup>In practice there is some tolerance for slight variability here. Fixed effects are not estimated without error.

<sup>2</sup>The phases of research described in Institute of Education Sciences solicitations map quite closely with the phases described here. For example, Design Studies can be interpreted as Phase I studies.

<sup>3</sup>Medical researchers distinguish efficacy from effectiveness. When the efficacy trials are conducted, every part of the study is generated with the highest quality. For example, the drug (treatment) is carefully produced in a laboratory setting to exacting standards. This contrasts with effectiveness studies, where the same drugs are generated at the factory level, and although production occurs to high standards, the standards are at a lower, more realistic level. Medical researchers want to know if

the treatment is efficacious before they decide to conduct effectiveness research. In some ways, a small study where the writer of the new curriculum teaches the experimental class would serve as a parallel example.

<sup>4</sup>*Dosing* is probably the wrong choice of word. Unlike doctors, teachers do not dose students. Curricula are delivered to groups of 25 to 30 students at a time. Not all students come ready for class. Some may be overprepared and underchallenged. I use the word only to provide consistency in argumentation. However, the issue is critical: It is more challenging in educational research and has not been acknowledged, let alone discussed, in our literature with any real depth in the context of randomized controlled trials.

<sup>5</sup>The treatment that is available to participants assigned to the control group is generally well specified. Normally, it is the drug that currently is deemed to provide the best current level of intervention efficacy.

<sup>6</sup>She remains nameless to protect the innocent.

## REFERENCES

- American Statistical Association. (2007). *Using statistics effectively in mathematics education research*. Washington, DC: Author.
- Bloom, B. S. (1972). Innocence in education. *School Review, 80*, 332–352.
- Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Educational Evaluation and Policy Analysis, 29*(1), 30–59.
- Brennan R. L. (Ed.). (2006). *Educational measurement*. Westport, CT: Praeger.
- Cronbach, L. J. (1976). *Research on classrooms and schools: Formulation of questions, design, and analysis* (Occasional paper). Stanford, CA: Stanford Evaluation Consortium.
- Education Sciences Reform Act, H. R. 3801, 107th Cong. (2002).
- Eisenhart, C. (1947). The assumptions underlying the analysis of variance. *Biometrics, 3*, 1–21.
- Fisher, R. A. (1937). *The design of experiments* (2nd ed.). Edinburgh, UK: Oliver and Boyd.
- Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trails in education. *Educational Evaluation and Policy Analysis, 29*(1), 30–59.
- 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.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.
- Raudenbush, S. W., Martinez, A., & Spybook, J. (2007). Strategies for improving precision in group-randomized experiments. *Educational Evaluation and Policy Analysis, 29*(1), 5–29.
- Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher, 31*(7), 15–21.
- Slavin, R. E. (2008). What works? Issues in synthesizing educational program evaluations. *Educational Researcher, 37*, 5–14.
- Sloane, F. C. (2005). *What can mathematics education learn from medical research?* Invited address to the Center for Mathematics Education at the Department of Curriculum and Instruction, College of Education, University of Maryland, College Park.
- Walker, D. F., & Schaffarzick, J. (1974). Comparing curricula. *Review of Educational Research, 74*, 83–111

## AUTHOR

FINBARR SLOANE is an associate professor in the Mary Lou Fulton College of Education, Division of Curriculum and Instruction, Arizona State University; [Finbarr.Sloane@asu.edu](mailto:Finbarr.Sloane@asu.edu). His research focuses on the learning of mathematics, methodology, and the modeling of student mathematical development in multilevel contexts.

Manuscript received January 10, 2008  
Accepted January 14, 2008