



# Randomized Trials in Mathematics Education: Recalibrating the Proposed High Watermark

Finbarr C. Sloane

The author reviews the recommendations in *Foundations for Success: The Final Report of the National Mathematics Advisory Panel (2008)* and agrees that a rebalancing of mathematics education research is timely and necessary, but questions whether randomized trials of small experimental studies and large field studies, without a clearer framing of the needed continuum of studies, can adequately rebalance the portfolio and address the Panel's "what works" questions. He offers one listing of the possible phases of research required to support high-quality causal inference in mathematics education as a way to foster continued debate about the ease of moving a model of research that works in one domain (drug trials) into the forced service of another intellectual domain (education).

**Keywords:** efficacy and effectiveness trials; internal and educational research; randomized clinical trials

The president of the United States established the National Mathematics Advisory Panel (NMAP) via Executive Order 13398 in April of 2006. He did this for a number of important reasons, the most critical being

to help keep America competitive, support American talent and creativity, encourage innovation throughout the American economy, and help State, local, territorial, and tribal governments give the Nation's children and youth the education they need to succeed. . . . [It] shall be the policy of the United States to foster greater knowledge of and improved performance in mathematics among American students. (p. 20519)

Following this expressed need, the goals of the Panel were quite broad and sweeping. These goals included the opportunity to make evidence-based statements about how and what mathematics should be taught in schools and about the way mathematics should be taught and what materials and curricula should be used, along with how teachers should be trained and supported professionally over the course of their careers. In addition, the Panel was given the task of and addressed questions regarding the needs for future research and, in effect, the quality of this research. In addressing this latter call for its input, the Panel argued that if the country is

to produce a steady supply of high-quality research that is relevant to classroom instruction, national capacity must be increased: More researchers in the field of mathematics education must be prepared, venues for research must be made accessible, and a pipeline of research must be funded that extends from the basic science of learning, to the rigorous development of materials and interventions to help improve learning, to field studies in classrooms. The most important criterion for this research is scientific rigor, ensuring trustworthy knowledge in areas of national need. (NMAP, 2008, p. 65)

Scientific research is seen in the Panel's mourning the current lack of randomized trials in mathematics education and its call for such trials. I am in agreement with this perceived need and the carefully crafted call, particularly as it pertains to and enhances the possibility of making qualified causal claims about research in mathematics education. However, I, like many others, see the need for research that more fully addresses the external validity of such causal claims (Briggs, 2008; Confrey, 2007; Cronbach, 1982; Shadish, Cook, & Campbell, 2002; Sloane, 2008) and research that is multilevel in nature (Raudenbush, 2008; Sloane, 2008). Consequently, the focus of this commentary is not to necessarily criticize the NMAP for its call for more randomized trials in mathematics education. Rather, I draw on research regarding the "rigorous scientific model" implied by the NMAP (i.e., the clinical drug trials model used in cancer research and other areas of medicine) to briefly discuss the central tenets of efficacy and effectiveness trials as they are related to education research. Implicitly, I criticize the Panel for assuming that issues of external validity can be easily addressed by simply generating studies that are large in size. I offer one possible framework for the long-term study of mathematics education as a research domain to generate debate about these issues. I do this to show that the model implicit in the Panel's recommendations is currently incomplete (and perhaps naive).<sup>1</sup> As such, I propose a recalibration of the high watermark for mathematics education research implicit in the many reports of the NMAP, one that questions whether we can simply force a particular model of medical research into the service of education research.

## Examining the NMAP's Recommendations Regarding Causal Research and Research Quality

Of the 45 recommendations made by the NMAP, my concern here is with recommendation 39. The NMAP noted in recommendation 39 that

Educational Researcher, Vol. 37, No. 9, pp. 624–630  
DOI: 10.3102/0013189X08328879  
© 2008 AERA. <http://er.aera.net>

it is essential to produce methodologically rigorous scientific research in crucial areas of national need, such as the teaching and learning of mathematics. . . . Although the number of such studies has grown in recent years due to changes in policies and priorities at federal agencies, these studies are only beginning to yield findings, and their number remains comparatively small. (NMAP, 2008, p. xxvi)

The emphasis on causal research is explicit throughout the Panel's many reports. The Panel explicitly noted that "the dearth of relevant rigorous research in the field is a concern" (NMAP, 2008, p. 63). By "rigorous" the Panel means experimental research, that is, research that explicitly follows the clinical trials model in medicine in allowing the researcher to make a particular type of causal claim. The Panel is clear in this call for causal knowledge, noting that this type of knowledge is essential to the production and evaluation of scientific research crucial to the current national need in the area of mathematics learning and teaching. The Panel noted that to produce such research will require enormous resources and that the "rigor and scale of the federal government's infrastructure for educational research must be dramatically increased" (NMAP, 2008, p. 63).

The Panel added that the nation's portfolio of research needs to be rebalanced and diversified to increase the amount and impact of experimental research. It argued that this research needs to be targeted along a continuum of studies from smaller scale experiments to larger scale field studies. For this to occur sensibly, and produce the needed quantity of experimental studies with the necessary quality, the NMAP went on to describe the needed changes in the current research infrastructure. The call for infrastructure change includes the adequate supply and appropriate training of researchers in this experimental paradigm, along with access to a large, diverse, and sufficient supply of schools and teachers who are willing participants in these studies. The Panel highlighted that these participants should have the "time, resources, and motivation to be partners in the research" (p. 64) and that they be able to use research findings in their decision making. Among other recommendations, the Panel noted the need for interdisciplinary researchers from a variety of disciplines (e.g., psychology, sociology, etc.) and for expedited human subjects' protection procedures. The Panel then underscored the need for a sufficient and stable source of research funding so that the highest quality research and training can be conducted.

The NMAP consistently noted the president's emphasis on the "best available scientific evidence," and as such, the Panel affords randomized trials in mathematics education research the highest standing. However, the Panel failed to distinguish between efficacy and effectiveness trials, and as such tended to confound the two in its call for more experimental research. In the next section, I articulate their differences in an effort to broaden and recalibrate the Panel's call for high-quality research in support of mathematics education. Furthermore, I elaborate some of the complications associated with the multilevel theorizing and modeling of research on instruction.

### **Locating Efficacy and Effectiveness Trials in Experimental Research to Support Mathematics Education**

I now describe the experimental methodology called for by the NMAP and draw attention to the critical need for education

researchers following this model to distinguish efficacy research from effectiveness research in mathematics education. As articulated earlier, I believe the chosen NMAP methodology to be that of randomized control trials (RCT), which are associated in particular with drug trials. The full RCT model engages a large and well-articulated number of phases. I briefly outline them here. For the four formal phases of the RCT model to be viable, a prephase of basic research must already exist. Phase I studies are used to establish feasibility; Phase II studies are used to demonstrate initial efficacy and to demonstrate improvement over historical norms; and Phase III studies are used to confirm efficacy by use of comparative randomized trials. Finally, Phase IV studies require follow-up studies in the real world of scaling to ensure effectiveness (American Statistical Association, 2007). Of note is that medical trialists distinguish between two types of research trials in the phases of their work: efficacy trials and effectiveness trials.<sup>2</sup>

*Efficacy trials* assess the value or worth of an educational treatment or program. To do so, these trials must provide tests of whether the treatment, instructional strategy, or curricular program does more harm than good when delivered under *optimal* conditions! In medical research, optimal conditions would include the highest quality production of the drug (or drugs) under study. Only later are such drugs produced under normal manufacturing conditions. In mathematics education, efficacy might include a selection of the most qualified teachers. (For example, those teachers originally trained by the curriculum innovator are assigned at random to treatment and control groups, ensuring, as best as one can, that the treatment has the highest possibility of being implemented with fidelity.) By way of contrast, *effectiveness trials* provide tests of whether the formally tested and efficacious treatment does more harm than good when it is delivered under real-world conditions.

Efficacy trials are considered a necessary condition for effectiveness trials to be considered. They are not sufficient for stating "what works." To reiterate, in the case of mathematics education, if the treatment cannot work under the most favorable conditions, it is unlikely to work in regular classrooms. However, just because the particular treatment works under the best conditions does not ensure that the same treatment will work in the same manner under more general, practical conditions.

The contrast between efficacy and effectiveness trials and the roles each play in knowledge generation and knowledge building is critically important to medical research and should be more important to education research. Efficacy and effectiveness trials are often conflated in education where the researcher assumes efficacy as a given. Put differently, rarely do we see efficacy research as a precursor to effectiveness research in education. These differences are not addressed in the NMAP report. In sum, it makes little sense to waste precious resources on effectiveness trials for treatments that have not been found efficacious.

In contrast, the shift from efficacy to effectiveness trials is explicitly acknowledged in the research models set out by two different divisions of the U.S. National Institutes for Health: the National Cancer Institute and the National Heart Lung and Blood Institute. These models are documented below, and links to the American Statistical Association's position on the RCT model are highlighted parenthetically. Both models support a

five-phase continuum of research; here the pre-Phase I stage, devoted to basic research, is considered a formal phase of the research program.

The National Cancer Institute cancer control research phases include

1. Hypothesis development (Phase I).
2. Methods development to ensure that accurate and valid procedures are available before a study actually begins (Phase II).
3. Controlled intervention trials (Phase III), where hypotheses developed in Phase I are then investigated with methodology validated in Phase II. Often the case control methodology is employed at Phase III.<sup>3</sup>
4. Defined population studies (Phase IV) to measure the efficacy of an intervention in a sizable, distinct, and well-described population.
5. Demonstration and implementation studies (Phase V).

The National Heart Lung and Blood Institute research spectrum includes the following phases:

1. Basic research (Phase I): Research that seeks new knowledge about normal and abnormal function of the heart, lungs, and blood and the etiology of their diseases.
2. Applied research and development (Phase II): Research that asks what new ways the results found in Phase II can be used in to achieve practical goals.
3. Clinical trials (Phase III): Trials conducted with large samples drawn from well-specified populations to determine the efficacy and safety of the interventions.
4. Prototype studies (Phase IV): Small-scale tests of refined programs using components of Phase III research to be efficacious; further development of methods for future research are also conducted here.
5. Demonstration and education research (Phase V): Tests of the effectiveness of the interventions.

What becomes clear when we look carefully at these research phases is that, along with distinguishing between efficacy and effectiveness trials, the phases are fundamentally nested in nature. That is, research at one level builds on that at another level or phase. This is a critical insight, as it suggests the possibility that the basic redress called for by the NMAP to add experimental studies, small and large, along some continuum of research is an overly simple response to the need to organize mathematics education research in support of improved learning on the part of students and teachers.

The Panel calls for research in support of the learning of specific mathematical content, that is, algebra; as such, the Panel calls for causal knowledge about “1) effective instructional practices and materials, 2) mechanisms of learning, 3) ways to enhance teachers’ effectiveness, including teacher education that focuses on learning processes and outcomes, and 4) item and test features that improve the assessment of mathematical knowledge” (NMAP, 2008, p. XXVI). The Panel falls short of fleshing out a theoretical map or framework for the conduct of such research.

Such a mapping, if nested appropriately, would optimize the possibility for research knowledge to build and accumulate more consistently over time. The expressed goal of this article is to offer one such conceptualization, for without such a framing the proposed goals for an improved research infrastructure in mathematics education are unlikely to be attainable. Such a framework (or body of work) must overtly acknowledge the need for research that can and does address both internal and external validity in the nation’s research portfolio (Briggs, 2008; Shadish et al., 2002; Sloane, 2008). In the terms of the medical research models described here, this portfolio of work must include efficacy trials as well as effectiveness trials before definitive statements can be made about “what works” in the very real and changing world of U.S. schools.

### **Needed: A Working Model for Causal Research in Mathematics Education**

No single article can suffice to fix the Panel’s perceived problem of knowledge accumulation in mathematics education. In an effort to generate much needed and open debate, I offer a simple, and in many ways a simplistic, *working model*, in the spirit of George Box’s much quoted insight that “essentially, all models are wrong, but some are useful” (Box & Draper, 1987, p. 424).

#### *Setting the Comparative Context*

A pharmaceutical company tests hundreds of new medications in trying to find one that will be both safe and superior to the standard treatment for a specific disease. However, individuals, although roughly “closed systems,” vary in their responses to this medication. So how are these pharmaceutical researchers to move forward?<sup>4</sup> Testing is now invoked and conducted in phases (or stages) to assess which medications add value to people’s lives. Most medications are eliminated at the initial stage, based on employing a small number of participants. If a medication looks promising, it is reexamined at a later phase of research where a more elaborate and severe test of its efficacy is made. The central problems then are to (a) generate an appropriate set of hierarchical phases for research and (b) carefully delineate the sizes and severities of experiments at the successive phases. This is done so that new medications considered “good” are unlikely to be discarded. It also serves to ensure that “poorer” medications do not receive expensive and resource intensive investigations. Without a shared set of research phases in place, the task becomes impractical if not impossible. The medical model, rendered here, is by default a linear model for causal knowledge building. It is clear that this linearity of intellectual effort will generate many delays in the knowledge-building process in education. However, I do not suggest that this linearity is the only way to build causal knowledge in education. For the purposes of this exposition, I follow that simple linear path.<sup>5</sup>

In mathematics education we are similarly interested in developing high-quality interventions and testing these interventions so that they can be either discarded or deployed into classrooms and schools. In contrast to drug trials, no clear and shared set of hierarchical phases currently exists to support this decision-making process for mathematics education researchers. The Panel works around this lack of direction by noting that “both smaller-scale

experiments on the basic science of learning and larger scale randomized experiments examining effective classroom practices are needed to ensure the coherent growth of research addressing important questions in mathematics education” (NMAP, 2008, p. 63).

For discussion purposes I propose a continuum with 10 phases of research, noting that without such an organizing structure it is unlikely that the Panel’s goals will be met. Because of space limitations I do not describe these phases completely, nor do I provide a worked example. The phases presented are hierarchical in nature and serve to stimulate much needed discussion. The proposed phases illuminate the differences between efficacy trials and effectiveness trials. In addition, multilevel issues are noted, when salient. In one case in particular (Phase V prototyping), I highlight the shift from cognitive or single classroom studies to instructional or multiple classroom studies. The purpose is to illuminate the hidden, but critical, inferential issues we need to negotiate in moving from analyses that focus on students as the analytic unit to analyses that consider the multilevel nature of schooling. As Raudenbush (2008) points out, we should not simply borrow inferential models from medical research. He argues that research involving students nested within classrooms must meet the “stable unit treatment value assumption” (SUTVA) as it manifests itself in educational settings before causal claims can be made. To do so, research on instruction must also meet additional assumptions of “no interference between classrooms” and “classrooms must stay intact” over the course of the study. Moreover, inferences can then be made only to the average teacher (or classroom) and not to the average student. These difficulties have been alluded to by Cronbach (1991) and by Cronbach and Webb (1976) in their reanalyses of extant data sets. Cronbach showed that results at one level of analysis may not hold up to scrutiny when examined from a multilevel perspective, a result that Slavin (2008) feels comfortable ignoring. It will not be as easy as the Panel assumes to move from single-level research to multilevel research or to generate crossover research without the development of multilevel theory that parallels the growth of multilevel modeling in the past 25 years.

### Ten Phases: A Very Brief Description

The 10 phases are presented in Table 1 with the following row headings: basic research, hypothesis development and measurement, pilot-applied research, Prototyping A, Prototyping B, efficacy trials, effectiveness trials, implementation trials, sustainability research, and scaling studies. For each proposed phase of research, I very briefly describe the types of research questions being investigated, the methods likely to be used, the samples and sampling plans, the nesting of samples, and their implications for analysis (see Table 1).

#### *Phase I: Basic Research*

Basic research as described in Table 1 is the bedrock of scientific investigation across disciplines. In fact, more rigorous research designs and complex theories are often (if not always) based on it. The researcher’s or practitioner’s intuition leads him or her to some conclusion based on limited data, with myriad alternative hypothesis and with great opportunities to be wrong. The point, however, is that the researcher also has the possibility to be right;

in either case the investigation provides the intellectual fodder, in the form of hypotheses, for more rigorous inquiries.

#### *Phase II: Hypothesis Development and Measurement*

Phase II research involves developing the hypotheses of Phase I by attaching to those early hypotheses very general conditions for success (design research), measures with which to gauge the complexity of the hypotheses and conditions under which they hold, and methods by which to scale those hypotheses if they are found useful. Phase II studies are conducted by many researchers, requiring a large amount of time to develop appropriate parameters, measures, and methods to study the hypotheses with greater clarity, simplicity, and rigor.

#### *Phase III: Pilot-Applied Research*

Phase III tests the refined hunches of Phase I, distilled into the hypotheses of Phase II, on small samples. Here researchers seek to find if the results found in Phase II hold under mildly experimental conditions. That is, are the hypotheses sufficiently developed to be tested in more precise ways?

#### *Phase IV: Prototyping A, and Phase V: Prototyping B*

I cluster these two phases because, although not the same, they are more interrelated than many of the other phases. Both involve small-scale tests of Phase III hypotheses that have been vetted against intuition (Phase I) and the development of measures, methods (Phase II), and basic experimental conditions (Phase III).

The fundamental difference between the two phases lies in the type of insight that is sought. In Phase IV, researchers seek insights about individual students, using them as the unit of analysis. Phase V, however, involves seeking insights at the classroom level while also modeling cross-level interactions between individual students clustered by classroom and teacher. The combination of Phases IV and V, then, provides researchers an opportunity to investigate how a treatment, once appropriately measured and hypothesized, affects a sample across its naturally occurring levels (Raudenbush & Bryk, 2002). As noted earlier, this work involves a critical shift in research trajectory from single-level analysis to multilevel analysis. This shift requires the development of pertinent multilevel theories in support of such investigation (Cohen, Raudenbush, & Ball, 2003; Sloane, 2005).

The inferences involved in the shift from studies of individuals to studies of groups, and the interactions between groups and individuals, are quite difficult to negotiate. As Cronbach (1991) noted, “Even investigators aware of the multilevel problem seem to have preferred the conventional individual-level analysis because its spuriously numerous degrees of freedom tend to make relationships look ‘significant.’ Multilevel reanalysis typically invalidates such conditions” (p. 397). In sum, the shift from analyses that look at learning to analyses that focus attention on instruction (in support of that learning) will require an intellectual leap of faith on the part of the research community at this phase for real progress to be gained. Drawing on Raudenbush’s (2008) interpretation of SUTVA, the basic classroom study involving two teachers, where one teacher is selected at random to deliver a treatment, does not afford us as a research community the opportunity to infer much about that treatment’s efficacy.

**Table 1**  
**Suggested Research Phases for the Development of Mathematics Education Interventions**

Phase	Overview	Methods, Sample, and Analytic Unit
I—Basic research	Discipline-based research (e.g., cognition in math)	Defined by the discipline
II—Hypothesis development and measurement	Develop hypotheses about possible approaches to promote learning based on models drawn from Phase I findings; develop the measures and methods with which to test hypotheses	Reviews, meta-analytic summaries, design research, exploratory research
III—Pilot-applied research	Preliminary tests of new approaches to achieve specific learning goals or short-term effects	Design research, quasi-experiments, experiments; sample size is small; the individual serves as the analytic unit
IV—Prototyping A	Small-scale tests of a refined program of instruction suggested by Phase III research	Design research, experiments, and quasi-experiments; sample size is small; the analytic unit is the student
V—Prototyping B	Small-scale tests of the refined program of instruction suggested in Prototyping A studies; building multilevel insight	Experiments and quasi-experiments; the classroom is the analytic unit; sample size is still small
VI—Efficacy trials	Randomized trials to determine efficacy of the intervention (e.g., teacher PD and new curriculum), as suggested to be effective by early phases of research	Randomized control trials with teacher as the unit; multilevel investigation with subpopulation analyses
VII—Effectiveness trials	Trials to determine the effectiveness of an intervention (shown to be efficacious) on a broader population	Large-scale experiments or quasi-experiments in real-world settings where delivery of the treatment is standardized and then carefully assessed
VIII—Implementation trials	Trials to determine the effectiveness of the efficacious intervention under real-world conditions	Large-scale experiments or quasi-experiments in real-world settings; delivery can vary naturally, or planned comparisons can be conducted; care must be exercised to assess the variation; nesting should be accounted for in the analyses; checks need to be made for fixed and random effects
IX—Scaling studies	Studies that determine the effects of an efficacious program when implemented in whole systems (e.g., school districts); explicit focus on depth, spread, and shift of program (see Coburn, 2003)	Large-scale, quasi-experiments with program when implemented in whole systems (naturalistic components in real-world settings; delivery will vary naturally; diffusion patterns need to be studied)
X—Sustainability research	Studies to examine the sustainability of the tested program in real-world settings to see what conditions need to be in place for the program to self-sustain	Large-scale, naturalistic evaluations; mortality of treatment and of program, and long-term effect on teachers should be examined; diffusion patterns need to be analyzed

### *Phase VI: Efficacy Trials*

An intervention or hypothesis under investigation at Phase VI has been carefully measured, translated into methodologies and analytic methods, and now requires *prime facie* evidence based on theory (a multilevel theory that acknowledges the linkages between instruction and learning). One does this with an RCT to eliminate potential threats to internal validity. Of note is that this phase does not guarantee external validity, which is saved for Phase VII research.

### *Phase VII: Effectiveness Trials*

The efficacious treatment rising from Phase VI research still may not work in the real world. In fact, the experimental act of

randomizing may not have accounted for a variety of naturally occurring problems simply because the act of randomizing was ultimately experimental. It does ensure an effect is there, yes; but it does not ensure that the effect persists when experimental conditions are removed. In essence, internal validity does not ensure external validity. Phase VII research seeks to test just that: Does an effect under experimental conditions (internal validity) persist in larger and more diverse populations under conditions that differ from those studied in the efficacy trial (external validity)?

### *Phase VIII: Implementation Trials*

An effective and efficacious treatment, one protected against the wiles of external and internal validity, now faces the real world in situ. This is when a treatment is given to larger populations and

allowed to bend to the variety of realities that individuals and groups thereof face on a daily basis. The myriad threats to internal and external validity, ruled out in the previous two phases, may find an alley in some previously unanticipated threat or event that requires researchers to return to an earlier point in the progression of phases. Effects may vary randomly or in a fixed manner across individuals and groups; interactions between groups and individuals may change the nature of the treatment or the interpretation of its effects. The variety of plausible problems is challenging to even the most closed mind, and ultimately they require that analyses in this phase of research be conducted with great respect for the complexity of students, classrooms, and the social network within which they exist.

### *Phase IX: Scaling Studies*

Finally a treatment developed with checks against external and internal validity is now ready to be implemented in settings where treatment fidelity is not monitored. The treatment may now deteriorate or fail to have an estimable effect. Worse yet, the treatment may hurt students (although with carefully completed previous phases, this is unlikely). With scaling studies, researchers will be able to investigate, with large-scale quasi-experiments, the effect of a treatment as it diffuses throughout a population. These researchers will be able to vet the experimental or quasi-experimental findings of previous phases against the real-world result in application.

### *Phase X: Sustainability Research*

A treatment that works for students and teachers, a population in the real world, in all of its complexity is not out of the proverbial woods yet. Over time the effect of the treatment may dwindle or disappear, increase or intensify in neither sought after nor invited ways. What is required for a treatment to function over time? Does it ultimately harm some group of individuals, students, or teachers, as they grow older or as the treatment grows older? Phase X research seeks to investigate these questions. The usefulness of this phase of research is well established in medical research (see Berenson, 2007) but has not been addressed in education research. Sustainability is central to the efforts of the Panel, for it assumes quite naively that proven change will consistently and positively affect practice in schools.

## **Conclusion: Drawing Causal Inferences From Mathematics Education Research**

Efficacy trials are a necessary condition in helping policy makers address “what works” questions in education. However, they are far from sufficient in serving the needs of such researchers. It is also true that they are but one step in a very long program of needed education research that must include multilevel insight. The shift from single-level analysis to multilevel educational theory and analysis is a critical, but not an easily negotiated, one. It is my hope that the program of intellectual effort outlined here as a series of research phases generates the opportunity for debate in the community that advocates for randomized trials in education. This debate is needed to ensure that mathematics education researchers do not simply transfer (or copy) a research model that cannot fully answer the needs of those involved in studying the

effects of new curriculum and instructional methods, let alone the myriad other intellectual trajectories in education research.

## **NOTES**

I am currently in receipt of funding from the National Science Foundation (NSF 0616306). The views expressed here are mine and in no way represent the position of the NSF. I sincerely thank the two external reviewers whose editorial critique and commentaries were extremely helpful to this writer. I especially thank Anthony Kelly, the special issue editor, for his thoughtful and encouraging comments. I also thank Brandon Holding of the Mary Lou Fulton College of Education for his critical insight, support, and editorial assistance. Finally, a special thanks to David Berliner, Gene Glass, and Mary Lee Smith for their professional interest and support of my research.

<sup>1</sup>See, for example, Alex Berenson's (2007) *New York Times* exposé on the clinical trial results associated with the cholesterol-lowering drug Zetia.

<sup>2</sup>For a more complete outline of this model, see Sloane (2008). In that article, I addressed the potential matching between medical research and education research in general, paying particular attention to multilevel issues of educational treatment and dosage.

<sup>3</sup>*Case control* is a type of epidemiological study design. Case-control studies are used to identify factors that may contribute to a medical condition by comparing patients who have that condition (the “cases”) with patients who do not have the condition but who are otherwise similar (the “controls”).

<sup>4</sup>In medical research individuals are perceived “roughly” as closed systems, and functionally they are closed systems, when contrasted with schools and classrooms. Consequently, the multilevel nature of the settings in which schooling occurs must be considered when evaluating either the efficacy or the effectiveness of educational interventions.

<sup>5</sup>The interested reader is referred to Kelly, Lesh, and Baek (2008) for a nonlinear take on the development of robust treatments.

## **REFERENCES**

- American Statistical Association. (2007). *Using statistics effectively in mathematics education research*. Alexandria, VA: Author.
- Berenson, A. (2007). Data about Zetia risks were not fully revealed. *New York Times*. Retrieved October 1, 2008, from <http://www.nytimes.com/2007/12/21/business/21drug.html>
- Box, G. E. P. & Draper, N. R. (1987). *Empirical model-building and response surfaces*. New York: John Wiley.
- Briggs, D. (2008). Synthesizing causal inferences. *Educational Researcher*, 37, 15–22.
- Coburn, C. (2003). Rethinking scale: Moving beyond the numbers to deep and lasting change. *Educational Researcher*, 32(6), 3–12.
- Cohen, D. K., Raudenbush, S. W., & Ball, D. B. (2003). Resources, instruction, and research. *Educational Evaluation and Policy Analysis*, 25, 119–142.
- 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, 195–213.
- Cronbach, L. (1982). *Designing evaluations for educational and social programs*. San Francisco: Jossey-Bass.
- Cronbach, L. (1991). Methodological studies: A personal retrospective. In R. E. Snow & D. E. Wiley (Eds.), *Improving inquiry in social science: A volume in honor of Lee J. Cronbach* (pp. 385–400). Hillsdale, NJ: Lawrence Erlbaum.
- Cronbach, L., & Webb, N. (1976). Between-class and within-class effects in a reported aptitude X treatment interaction: Reanalysis of a study by G. L. Anderson. *Journal of Educational Psychology*, 76, 717–727.

- Executive Order 13398. (2006). *Federal Register*, 71(77), 20519–20520.
- Kelly, A., E., Lesh, R. A., & Baek, J. Y. (2008). *Handbook of design research methods in education: Innovations in science, technology, engineering, and mathematics learning and teaching*. New York: Routledge, Taylor and Francis Group.
- National Mathematics Advisory Panel. (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel*. Washington, DC: U.S. Department of Education.
- Raudenbush, S. (2008). Advancing educational policy by advancing research on instruction. *American Educational Research Journal*, 45(1), 206–230.
- Raudenbush, S., & Bryk, A. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.
- Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.
- Slavin, R. (2008). What works? Issues in synthesizing educational program evaluations. *Educational Researcher*, 37, 5–14.
- Sloane, F. (2005). The scaling of reading interventions: Building multi-level insight. *Reading Research Quarterly*, 40, 361–366.
- Sloane, F. (2008). Through the looking glass: Experiments, quasi-experiments, and the medical model. *Educational Researcher*, 37, 41–46.

#### AUTHOR

FINBARR C. SLOANE is an associate professor in the Mary Lou Fulton College of Education, Division of Curriculum and Instruction, Arizona State University, Room 203b, Farmer Building, Tempe, AZ 85287; *Finbarr.Sloane@asu.edu*. His research focuses on the learning of mathematics, behavioral methodology, and the modeling of student mathematical development in multilevel contexts.

Manuscript received July 26, 2008

Revision received October 20, 2008

Accepted October 20, 2008