On Scientific Research in Education: Questions, Not Methods, Should Drive the Enterprise

Richard J. Shavelson[1]
Stanford University
and
Lisa Towne
National Research Council

For over 100 years, education research as a scientific endeavor has been at the center of scholarly and political debate. With the recent advent of "evidence-based" policy and practice in education and related fields the debate has taken on heightened importance and political overtones. Indeed, in the summer of 2000 a bill to reauthorize the primary federal education research agency included a legislatively, not scientifically, devised definition of what constitutes "scientifically based research" in education. This action signaled the field's lack of credibility with policymakers and the high stakes associated with articulating and upholding standards of high quality science.

This is the context, in the winter of 2001, when a National Research Council committee[2] met to address three related questions in response to a request by the National Educational Research Policy and Priorities Board:[3] (1) What are the principles of scientific quality in education research? (2) How can a federal research agency promote and protect scientific quality in the education research it supports? And (3) how can

[1] Based on remarks given at a workshop of the National Research Council Committee on Research in Education. For more information on the event and the committee's work, see http://www7.nationalacademies.org/core/.

[2]Donald I. Barfield, Robert F. Boruch, Jere Confrey, Rudolph Crew, Robert L. DeHaan, Margaret Eisenhart, Jack McFarlin Fletcher, Eugene E. Garcia, Norman Hackerman, Eric Hanushek, Robert Hauser, Paul W. Holland, Ellen Condliffe Lagemann, Denis C. Phillips, and Carol Weiss. Lisa Towne served as study director and I chaired the committee.

[3] NERPPB was the policy arm of the former U.S. Office of Educational Research and Improvement, which was replaced by the Institute of Education Sciences with the passage of the Education Sciences Reform Act of 2002.

research-based knowledge in education accumulate? About a year later, we published our answers to these questions (Shavelson & Towne, 2002; http://books.nap.edu/catalog/10236.html), being careful to point out that our charge was to explore the *scientific basis* of education research, and that "…historical, philosophical, and literary scholarship can and should inform important questions of purpose and direction in education" (p. 26).

Since the release of the NRC report, public debate has intensified, and several congressional and executive branch actions have focused on bringing scientific research to bear on education policy and practice. A strong focus of these efforts has been on shoring up the (perceived) low quality of current scholarship by pushing the use of randomized field trials—held up as the "gold" standard—in education research.

Research Questions and Methods

In the course of our deliberations, we inevitably took up this contentious topic, the design of education research. At one extreme we found some experimentalists (and policy makers) who believed that unless research involved a randomized trial, it was not scientific and not worth doing. At the other extreme were postmodernists who didn't put much stock in scientific research of any kind. Committee members held a wide range of views on what they personally considered to be scientific.

Perhaps the Committee' greatest contribution was to recognize that defining scientific research by method was wrong minded. It's the question—not the method—that should drive the design of education research or any other scientific research. That is, investigators ought to design a study to answer the question that they think is the important question, not fit the question to a convenient or popular design.

Incidentally, once this insight had been reached, unbeknownst to my colleagues, a ditty from Gilbert and Sullivan's *Mikado*[4] started running through my mind,

> "His object all sublime
> He shall achieve in time--
> To let the punishment fit the crime--
> The punishment fit the crime."

I simply substituted question for crime and method for punishment:

> "His object all sublime
> He shall achieve in time--
> To let the method fit the question--
> The method fit the question."

In hindsight, this idea should have been obvious once we had argued our way to the conclusion that scientific research in education was, in general, like scientific research in the social and natural sciences, and should:

- Pose significant questions that can be investigated empirically
- Link research to relevant theory
- Use methods that permit direct investigation of question
- Provide a coherent, explicit chain of reasoning to rule out counter-interpretations
- Replicate and generalize findings across studies
- Disclose research to encourage professional scrutiny and critique

To be sure, each area of science has crafted its methods to fit its paradigmatic questions and phenomena, something that Thomas Kuhn pointed out decades ago. This is true of education as well as the natural and social sciences. What is common among methods is that they build into research design the characteristics of scientific research,

[4] To hear the song and read (sing?) the lyrics, go to ttp://math.boisestate.edu/gas/mikado/webopera/song17.html

such as ruling out counter interpretations and generalizability, enumerated above.

Types of Questions and Corresponding Methods

The insight that the method should follow from the question, now pretty obvious and something graduate students often hear from their professors, led to another conundrum. There are so many research questions, how do we say anything cogent about method? The Committee reasoned that most scientific research questions were of three general types: (1) What's happening? (2) Is there a systematic (causal) effect? And (3) what is the causal mechanism or how does it work?

What's Happening? The question—what's happening?—asks for a description. We could ask this question in a materials science laboratory or in a middle school. We could describe the years of experience possessed by elementary school teachers in the U.S., or the types of science instruction students receive, or the changes in students' mathematics achievement over 20 years. In these cases, statistical estimates of population parameters could be obtained from available national surveys. Or we could describe what school, home and community look like though the eyes of an inner-city youth using ethnographic methods. Or we could describe different approaches to the assessment of learning in colleges and universities that have been nominated as "exemplary" using case study methods.

Holland and Eisenhart's (1990) study exemplifies scientific research into what's happening, and, as we shall see, beyond. They were concerned about explanations for why so few women pursued careers in non-traditional majors such as science: women were not well prepared before coming to college, women were discriminated against in college, women did not want to compete with men for jobs.

They began their study by *describing, in depth over several years* through *ethnography*, the lives of 23 volunteer women from two small public residential colleges. Half of these women pursued traditional careers and the other half non-traditional careers. They were matched on grades, college majors, college activities, and college peers. Based on extensive, detailed data collected through participant observation and interviews, Holland and Eisenhart found that what best described these women's academic pursuits were, *contrary to popular conjecture*, how they viewed the value of schoolwork, what their reasons were for doing school work, and how they perceived financial and opportunity costs. Simply put, detailed description of the college lives of these women portrayed their career trajectories in a very different light than the distal conjectures based on statistical data.

Now if you are thinking, "well, this is just idiosyncratic description and is suspect," Holland and Eisenhart one up you. They took the next step, going beyond description and entering the arena of model testing (see below) by predicting with their descriptive models what each of the 23 women would do after college: continue in school, get a job in her field, get a job outside her field, get married, etc. At the end of 4 years and another 3 years, they followed these women up with telephone interviews. In *all* 23 cases, predictions based on their model of commitment to schoolwork were confirmed. In all cases, their model provided a better predictor than did data on precollege preparation (grades and courses taken), discrimination against women, or feelings about competing with men.

<u>Is There a Systematic (Causal) Effect?</u> Questions about effects are, ultimately, questions about causal effects. Did *x* cause y? Perhaps the most widely known study of

systematic effects in education is the Tennessee randomized trial on class size reduction. The question posed by the Tennessee legislature was whether reduced class size would improve students' achievement (Finn & Achilles, 1990, 1999). To this end, within 79 schools across the state, a total of 11,600 students were randomly assigned to a regular class (22-26 students), a class with an aide (to decrease the student/adult ratio economically) or to reduced class size (under 13-17 students). Three findings stand out: (1) students in small classes outperformed students in the other classes, (2) minority students particularly benefited, and (3) the effect persisted when students returned to regular class sizes. Here the method, randomized trial, followed directly from the legislature's question and was feasible and ethical to implement. In such cases as this, randomized trials are the preferred method for ferreting out causal effects.

There are many cases, however, when randomization may not be feasible. Such cases include the effect of smoking on health and longevity, and the effects of hunger, alcohol use, drug use or child abuse on students' academic performance. For these research questions, other methods need to be used, are available, and include quasi-experiments (control and experimental groups without random assignment), correlational studies using large-scale probability-sampled data sets (that adjust for selectivity bias), and various time series designs. To be sure, as you move away from the randomization in some cases uncertainties increase; nevertheless, causal interpretations are possible and replication is important to increase confidence in the interpretations.

Loeb and Page's (2000) study of teacher salaries exemplifies the application of correlational (structural) modeling in a situation where random assignment is unlikely. They asked, "If teacher quality affects student achievement, why do studies that predict

student outcomes from teacher wages produce weak results?" That is, shouldn't teachers' salaries reflect, at least to some degree, teacher quality after controlling for other things?

Loeb and Page tested two competing models. One was the usual production function model that links inputs (salary) to student outcomes (dropouts in this case) after controlling for relevant variables. The second model followed their reasoning that there are other things in the lives of teachers than salary that may have meaning, and there also may be local job markets that provide attractive alternatives to teaching in the area. So their second, competing model incorporated opportunity costs into the production function: non-pecuniary rewards and competition in the local job market. They replicated prior research with the usual production-function model, showing a weak effect of salaries on outcomes. However, once they adjusted this model for opportunity costs (non-pecuniary and job market incentives), they found that raising wages by 10 percent reduced high school dropout rates by 3-4 percent.

Three points about studying causal effects seem appropriate here. First, in dealing with causal assertions we are always trying to rule out all the possible counter hypotheses that we know of at the time. As a research program moves along, new challenges (counter hypotheses) arise and get ruled out; in this way confidence increases in the causal interpretation. Oftentimes we don't know all the counter hypotheses; challenges arise with novel counter-interpretations, and research and debate continues as it has with the Tennessee study. This type of debate—hypothesis/counter-hypothesis—is the basis of science and should be looked upon positively and not as "backbiting" among scholars

with different views when the issue is one of interpretation; it is backbiting when personal attacks are made.

A second point has to do with the role of description in causal studies—what's happening? When feasible, descriptive research should be used in causal studies to help us understand, as fully as possible, what "treatments" were actually implemented, and to reveal what possible causal mechanisms might be operating.

And the third point is that establishing a causal effect may be necessary when possible but not sufficient in policy and practice. The questions of mechanism and context (see point 2) inevitably should arise in order to design education policies or practices (Cronbach, Ambron, et al., 1980). We need to understand how interventions were articulated and implemented in diverse contexts with whom, under what conditions with what resources in order to design more than superficial education policy.

What Is the Causal Mechanism or How Does It Work? The third type of research question focuses on the mechanism that creates a causal effect. For example, reducing class size seems to have a salutary effect according to the Tennessee study. But what was the mechanism that caused the effect and why did it persist even after students returned to regular class sizes (Grissmer, 1999)? Was the effect due to an increase in the number and personal nature of teacher-student contacts or to less off-task student behavior (Blatchford, 2003) or to the level of student engagement (Finn, Pannozzo, & Achilles (2003)?

Empirical studies of mechanism, following on studies that have established causal effects are most common. Bryk, Lee and Holland (1993) sought to understand the causal mechanism(s) underlying the causal evidence that Catholic schools outperform public

schools in the U.S. This longitudinal study used both qualitative (e.g., case studies of effective Catholic schools) and quantitative data to address the mechanism question. Three potentially explanatory models were tested: (1) sector effects only (spiritual and private characteristics of Catholic schools), (2) compositional effects (kinds of students attending Catholic schools), and (3) school effects (school operations contributing to school context). A combination of models, characterizing "…the *coherence* of school life in Catholic schools … most clearly accounts for it relative success in this area" (Shavelson & Towne, 2002, p. 119).

Nevertheless, there is another way to approach the question of mechanism—namely, to build an artifact based on a causal theory and establish its causal effect. Studies such as "design experiments" or "design studies" posit a theory with a causal mechanism, and design educational artifacts (e.g., a curriculum, a computer application) and iteratively test them out in complex real-world classroom environments, revising both artifact and theory along the way. Once evidence accumulates to suggest a causal mechanism, the onus, of course, is on design researchers to then establish generalizable causal effects (Shavelson, Phillips, Towne, & Feuer, 2003).

Concluding Comments

If I leave you with anything from the NRC Committee it is this: the research question matters; the research design has to follow the question. There are a lot of important questions in education for scientific researchers to address. These questions could be descriptive, they could be causal, or they could be about mechanism. In a program of research, all three types of questions should be included. In this case, a

program of research would embrace multiple methods, each fitting the particular descriptive, causal or mechanism question at hand.

If the goal is to estimate systematic effects—as is often the case in program evaluations and policy studies—logically randomized trials should be the preferred method if they are feasible and ethical to do. Where they are not, quasi-experimental, correlational, and time-series designs provide viable alternatives. Nevertheless, regardless of research design (experimental, quasi-experimental, correlational), and especially in selling the so-called "gold standard" (randomized trials), difficulties arise. One problem has to do with the fidelity of treatment implementation. You claim that treatment T caused the effect but in fact it turns out to be T*, which isn't quite the treatment that you thought it was. A second difficulty lies in the variability in treatment implementation. Observations of treatment implementation may reveal variability, from T to T* to T** to T*** (etc.); this variability needs to be captured and its causal effect understood. A third difficulty is that the control group may (1) perform in such a way as to it looks like the experimental treatment, something not uncommon in experiments involving teachers; or (2) may not provide a relevant contrast for policy purposes, for example, the teacher aide control condition in the Tennessee study was irrelevant for California's implementation of class-size reduction policy. A fourth difficulty is that the outcome measure (e.g., broad-band achievement test with a single dimension) is inadequate—it does not measure all of the things or even many of the things that are important—what psychometricians call construct under-representation. Moreover, the selection of the outcome measure might privilege one treatment over another at the outset of the study. Yet such limited measures are commonplace in policy research. And

finally, of course, there is always the issue of external validity—the extent to which the experimental treatment generalizes to real-world contexts. The trade off between randomized trials and in situ studies (e.g., quasi-experiments, corrleational) needs to be weighed against the credibility and generalizability of findings.

References

Blatchford, P. (2003). A systematic observational study of teachers' and pupils' behaviour in large and small classes. *Learning and Instruction, 40*(6), 569-595.

Bryk, A.S, Lee, V.A., & Holland, P.B. (1993). *Catholic schools and the common good*. Cambridge, MA: Harvard University Press.

Cronbach, L.J., Ambron, S.R., Dornbusch, S.M., Hess, R.D.,, Hornik, R.C., Phillips, D.C., Walker, D.F., & Weiner, S.S. (1980). *Toward reform of program evaluation*. San Francisco: Jossey Bass.

Finn, J.D., & Achilles, C.M. (1990). Answers and questions about class size: A statewide experiment. *American Educational Research Journal, 27*(3), 557-577.

Finn, J.D., & Achilles, C.M. (1999). Tennessee's class size study: Findings, Implications, Misconceptions. *Educational Evaluation and Policy Analysis, 21*(2), 97-109.

Finn, J.D., Pannozzo, G.M., & Achilles, C.M. (2003). The "why's" of class size: Student behavior in small classes. *Review of Educational Research, 73*(3), 321-368.

Grissmer, D. (1999). Class size effects: Assessing the evidence, its policy impoications and future research agenda. *Educational Evaluation and Policy Analysis, 21*(2), 241-248.

Holland, D.C., & Eisenhart, M.A. (1990). *Educated in romance: Women, achievement, and college culture*. Chicago: University of Chicago Press.

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.

Shavelson, R.J., & Towne, L. (Eds.) (2002). *Scientific research in education.* Washington, DC: National Academy Press.