Identifying Scientifically-Based Research in Education

Stephen W. Raudenbush
University of Michigan

Prepared for the Scientifically Based Research Seminar, February 6, 2002

Sponsored by Assistant Secretary Susan B. Neuman and the Office of Elementary and Secondary Education, U.S. Department of Education,

January 30, 2002

Identifying Scientifically-Based Research in Education

In May of 1999, the American Academy of Arts and Sciences hosted a conference on ways to improve the scientific quality of educational research. Among the organizers were two men who had played a central role in a similar project 40 years ago. Howard Hyatt and Frederick Mosteller's concern in the 1950s and 1960s was not the quality of research in education but rather the quality of research in medicine.

Hyatt and Mosteller argued in those days that carefully controlled clinical trials ought to become the norm for deciding which new vaccines, new surgical procedures, and new medications should be widely prescribed.

Their arguments met considerable skepticism. Hyatt told a story about a widely publicized debate between him and a heart surgeon. The question was whether it was ethical and feasible to conduct experiments in which heart patients would be assigned to a new surgical procedure versus a standard medical treatment. The heart surgeon asked: "Sir, have you ever held the beating heart of a human being in your hand?" The surgeon argued that the cold logic of science should not replace the clinical judgement of the seasoned practitioner.

In response, Hyatt and Mosteller noted that, in many cases, the profession really doesn't know what the best treatment is for a given disease. In that situation, it is unethical for us NOT to use the best available scientific methods, including experiments, to find out what works best. Once we know how best to deal with a given disease, many will benefit, revealing the true ethical character of the decision to conduct experiments.

Over the past 40 years, Hyatt and Mosteller's point of view has largely won out in the field of medicine. We now accept and admire the commitment of medical professionals to base their diagnoses and prescriptions on clinical trials in which patients are randomly assigned to alternative treatments.

The parallels between the debate in medicine then and the debate in education now are striking. At a recent conference, I recommended that our best ideas about how to improve teaching ought to be tested scientifically. A well-known educational researcher accused me of totalitarian thinking that unethically denies parents and teachers their rights.

People hold strong opinions about many important questions in education:

* Would a structured academic curriculum improve the pre-literacy skills of preschoolers? Would it harm their emotional development?

* What mix of methods in early reading instruction has the best long-term benefits for reading comprehension?

* Does math instruction based on the new NCTM standards boost students' mathematical reasoning?

* Does ending social promotion and increasing remedial instruction boost learning? Does it raise the drop-out rate?

* Can a voucher program boost the learning rates of children living in poverty?

Educators strongly disagree about these questions. We don't currently know the answers. The ethical action is not simply to stick to our personal beliefs on these issues but to do the much harder work of getting the needed empirical evidence.

My central contentions are two: first, we can answer questions like those posed above using scientific methods.

Second, the criteria we ought to use in evaluating studies designed to answer these questions are no different from the criteria used to judge scientific research in medicine.

What Caused the Change in Medicine?

It's instructive to ask what caused the sea-change in thinking about medical research over the past 50 years.

One of the most influential experiences concerned the effectiveness of the Salk vaccine for polio (Meier, 1972). Early studies compared those who received the vaccine to those who did not. The results were discouraging: people receiving the vaccine had polio rates that were as high as those who did not receive it. But there was a problem: Subsequent studies showed that high income families were more likely than low income families to receive the vaccine. Moreover, high income families were also at GREATER RISK of contracting polio. So the early studies were biased against finding a positive effect of the vaccine.

A subsequent large scale study in 1954 assigned persons at random to receive the vaccine versus a placebo. The results unmistakably supported the vaccine. Random assignment assured that the two groups had the same risk of contracting polio in the absence of the vaccine. The large difference in disease rates that emerged during the study could be plausibly explained only by one factor: access to the vaccine. The earlier poorly controlled studies had it wrong; the later well controlled study had it right. Since then, untold millions have benefited from ever-improved versions of the vaccine. Experimentation played a key role in this process.

Parallels Between Medicine and Education

The parallels in educational research are striking. The first widely-publicized evaluation of Head Start indicated that kids who had received Head Start had no better cognitive skills than kids who had not received Head Start. Many declared Head Start a failure. Subsequent investigation showed clearly, however, that the families of Head Start kids were, on average, poorer than the families of non-Head Start kids. In light of these higher poverty levels, one might have expected the Head Start kids to do significantly worse on the cognitive test than the non-Head start kids if Head Start had no effect. So some argued that the evaluation results showed a positive effect of Head Start. Unfortunately, the experiment that might have settled the issue was never conducted.

In the early Salk Vaccine studies and in the Head Start evaluation, the socioeconomic status of the families was what statisticians call a "confounding variable" or a "confounder" for short. A confounder is a pre-existing characteristic of the participants in a study that is related to the outcome and also predicts treatment group membership.

In the Salk vaccine case, family income was linked to the disease -- high-income kids were more likely to get polio -- and to treatment group membership: high-income kids were more likely than low income kids to get the vaccine. Family income was therefore a confounder. To ignore the effect of this confounder was to bias the study against finding an effect of the vaccine.

In the Head Start case, child poverty was negatively related to the cognitive outcome but positively related to membership in Head Start. Ignoring poverty biased the evaluation against finding a positive effect of Head Start.

The Power of Experimentation

One of the most common strategies in research is to try to identify and control for confounding variables. So in the Head Start study, one might match kids on family income and compare Head Start kids to the matched non-Head Start kids. This will eliminate family income as a confounder. The problem is that there are many potential confounders. We can't measure and control for all possible confounders.

Without random assignment, the burden is always on the researcher to show that relevant confounders were controlled. There is always some uncertainty that an important confounder was ignored, biasing the evaluation.

The power of the randomized experiment is that it controls all confounders. When kids -- or classrooms, or schools -- are randomly assigned to program A versus program B, we know that there are no confounders. Though the groups may still differ somewhat by chance on background characteristics, the differences are likely to be small. Moreover, our methods of statistical hypothesis testing accurately gauge the uncertainty that arises from such chance differences.

Questions and Answers About Scientific Research in Education

I am allotted a short time in this talk, yet many good questions follow from the discussion so far. Let me pose a few of the obvious questions and, in each case, provide my own view of the answers. In this way I hope to stimulate rather than end the important debate over scientific methods in educational research.

1. *Am I saying that only studies that use random assignment are scientific?*

No, I am not saying that, for three reasons.

First, random assignment is relevant only when *causal questions* are on the table. Many key questions in education are not causal. For example, we might ask:

* Have high school graduation rates changed over the past 10 years? Which kinds of kids, in which cities and states, are at highest risk of dropping out?

These are not causal questions but they do have scientifically-based answers.

Second, even when the question at hand is causal, it may be impossible to do a randomized study. Medical researchers have found a causal link between smoking and lung cancer without randomly assigning patients to smoke two packs a day. We need to know how family conflict affects school learning but we will never get the answer to that question from a randomized experiment.

Third, randomized experiments sometimes create artificial circumstances that limit the generalizability of findings.

2. *Ok, but suppose I do have a causal question. How do I judge the scientific quality of a study that does not use random assignment?*

Perhaps the key feature of scientific research is that the researcher is obligated to systematically and painstakingly evaluate alternative explanations for any finding of interest. Suppose we find that children who experience a new writing program display higher-quality writing than children who do not receive the program. We don't automatically conclude that the program is effective. Instead, we ask: Based on available theory and past evidence, what the likely confounders? Were children in the new writing program advantaged on those confounders?

A scientist is expected to search for disconfirming evidence. For example, perhaps the teachers in the new program were especially highly motivated. Maybe they simply spent more time teaching writing than did teachers not in the program.

A researcher might also ask: How does the writing program actually work? Which ingredients of that program are most likely linked to better writing? Were those components actually implemented?

If we can do a randomized experiment, we can eliminate many sources of bias. But the researcher is still obligated to consider alternative explanations for why the treatment did or didn't work. Even in a randomized experiment, critics may claim that the wrong outcome variables were measured or that the study results do not generalize to the population of kids of interest.

Moreover, randomized experiments are never perfectly implemented. Some schools or classes or kids will drop out of the treatment group and the control group, potentially producing subtle or not-so-subtle biases.

What makes a causal comparative study scientific, then, is not simply whether the investigator used random assignment. In every study, the investigators must critically evaluate competing explanations for what was found and why.

3. Isn't it a little polyannish to expect researchers to police themselves in this way? After all, researchers are human beings with biases.

The burden of objectivity does not fall entirely on the shoulders of the individual researcher. The role of the scientific community is key. A commitment to evaluate alternative explanations and to search for disconfirming evidence is what we call objectivity. While individual scientists are expected to uphold objectivity in their work, objectivity is, in the final analysis, a collective responsibility of the scientific community.

The methods of a study should be open to public scrutiny and data should be available for re-analysis. Findings should be subjected to rigorous peer review. And key conclusions emerge typically from convergent results over multiple studies conducted by multiple investigators whose personal viewpoints typically differ. A healthy scientific community is essential in examining the results from such streams of research.

Scientific evidence from a single study is rarely decisive. Instead, scientific knowledge emerges as a community of scientists evaluate a stream of studies over time -- more on this point later.

4. *Are randomized studies possible in education?*

They clearly are possible and often useful. We may point to the Tennessee class size experiment, which Frederick Mosteller has called the most important educational study in

decades. There have been randomized evaluations of whole school reform (Thomas Cook's studies of James Comer' program (Cook, et al., 1999a; Cook, Hunt, and Murphy, 1999b), and randomized studies of the Reading Recovery program. There are ongoing randomized studies of vouchers, of neighborhood effects on educational achievement, and many studies of violence prevention and drug prevention in school settings (Cook, 2001). Randomized experiments cannot answer every question but their use in education can certainly be expanded.

5. *How can a randomized experiment in education be done ethically?*

Consider a popular program such as Success for All, which now is working in more than 1000 elementary schools in an attempt to boost early literacy (Slavin, in press). Many schools want to adopt the program but it is expensive and the resources available are limited. Indeed, it is impossible to simultaneously implement the program in every school that wants it.

One might seek schools to volunteer to get the program at no cost or a reduced cost. All volunteering schools would ultimately receive the program, but the timing -- that is, which schools get the program first -- would be decided by a lottery. A lottery is a perfectly fair way to decide this question, given that resources do not allow all interested schools to receive the program simultaneously. The schools assigned to receive the program later become a randomized "wait list control group" whose outcomes can be compared to the outcomes of schools receiving the program during the waiting period.

Two strategies make this kind of approach ethically sound and practically feasible: 1) the use of a wait-list control group; and b) the assignment of schools rather than kids to treatments.

In other cases, for example, in the case of studying a tutoring program, assignment of kids at random to a treatment group or to a wait-list control will make good sense.

And in still other cases, there will be no true control group. Rather, there may be two alternative programs -- both attractive -- that can be compared. If we really don't know which works better, one can argue for randomized experimentation, providing, of course, that participants are willing to try either approach. This latter condition may not hold, in which case a well-controlled but non-randomized study may be needed.

6. *I mentioned that not all scientific question in education are causal. What are some examples*?

Over the past 30 years, the National Center for Education Statistics has commissioned a number of large-scale surveys. Thousands of scientific studies have used these data to help us understand:

* the levels of literacy and content knowledge of kids of varied background in varied states at varied times;

* how literacy levels and content knowledge are changing over time;

* how the mathematical and scientific understanding of US children compare to that of children in other countries;

* how approaches to teaching in math and science vary across schools within the US and between the US and other countries;

* how well qualified US secondary teachers are to teach their assigned content and where the shortages in teacher qualifications show up;

* the access of kids of varied background to various educational resources;

* which kids in which kinds of schools and communities are at highest risk of dropping out of school.

* how various kinds of schooling experience correlate with post-secondary educational opportunities and learning;

* how schools are financed and how school finances are linked to opportunities for learning;

* the levels of adult literacy in varied occupations and how this compares to literacy in other societies.

There are many other examples (c.f., Whiteley, Weinshenker, and Seelig, 2002). These studies provide vast and useful scientific evidence about conditions of US education and targets for improvement.

7. *How are these "non-causal" studies judged?*

We need to know in every case if the sample selected represents the population we are interested in. We need to know if the methods of asking questions (e.g., by interviewing, questionnaires, tests, or administrative data collection) produce reliable and valid indicators of the variables of interest. We need to know if the methods of analysis are accurate. We need to ask whether alternative explanations have been painstakingly assessed.

But there is no set of simple rules for judging the validity of scientific research. Instead, we must reply upon a community of experts to judge scientific claims through well-organized peer review.

8. *So far I have mentioned only quantitative research. Does qualitative research play a role in making educational research more scientific?*

Yes, without doubt. Qualitative research has provided:

* careful description of how the most expert primary school teachers teach (for example, how they teach fractions or beginning reading);

* how children of varied cultural backgrounds experience the transition from home to school;

* how differences between "school language" and "home language" shape children's participation in classroom discourse;

* vivid descriptions of how individual children learn.

There are many more examples. These studies give us new ideas about teaching, new insights about why programs work when they do work. Qualitative research can spur creativity in educational research by giving us compelling "up-close" descriptions of how teaching and learn work -- or don't work.

9. *How does one combine insights from various kinds of inquiry*?

Another analogy to medicine is perhaps instructive.

I mentioned earlier that public health scientists became convinced that smoking causes lung cancer even though it was impossible to test this link with randomized experiments.

First, a series of well-designed non-experimental studies showed that smokers were more likely than non-smokers to get lung cancer. Moreover, researchers found that, among smokers, the amount smoked and the probability of lung cancer were linked. As these studies controlled for more and more potential confounders, it became more and more difficult to claim that biases caused by unobserved confounders explained the correlation between smoking and lung cancer.

Second, it was possible to conduct randomized experiments on animals. Scientists knew that they could not automatically generalize these results to humans, but the results of these experiments on animals were consistent with the growing body of non-experimental evidence on humans, helping shift the burden of proof to those who denied the causal connection between smoking and lung cancer.

Third, careful examination of the lungs of smokers revealed that the kind of damage to their lung tissue was consistent with the causal hypothesis.

Thus, three kinds of studies contributed to the emerging scientific consensus: non-experiments (essentially surveys) comparing smokers and non-smokers; true experiments (on animals), and what might be called qualitative research -- careful inspection of lung tissue. The growing weight of evidence from this stream of research created a new consensus among scientists who had previously disagreed: smoking causes lung cancer.

Research evidence from varied studies is combined similarly in education. For example, despite the intense controversy over how to teach early reading, many points of consensus have emerged (Snow et al., 1998).

10. *The discussion so far conveys considerable enthusiasm about the role of science in education. Is there a risk in unrestrained enthusiasm?*

If science is to make a sustained contribution to education, we have to be careful not to oversell what science can do. Twice during the 20th century, educational researchers created overly-optimistic expectations for science (Raudenbush, 1982). When the results failed to meet these expectations, the scientific approach was discredited.

Consider, for example, E.L. Thorndike's lead essay in the founding issue of the Journal of Educational Psychology in 1910:

> "A complete science of psychology would tell every fact about everyone's intellect and character and behavior, would tell us the cause of every change in human nature, would tell us the result which every educational force -- every act of every person that changed any other or the agent himself -- would have. It would aid us to use human beings for the world's welfare with the same surety of the result that we now have when we use falling bodies or chemical elements. In proportion as we get such a science we shall become masters of our own souls as we are now masters of heat and light. Progress toward such a science is now being made." (Thorndike, 1910:8)

Thorndike's hopes for the role of education were unrealistic. The failure to meet these inflated expectations overshadowed very real but slow progress in the study of education. As a result, public interest in educational research declined. Much later, in the 1960s and 1970s, advocates of systematic evaluation of government anti-poverty programs again over-sold the power of science. The result was another cycle of disappointment and retreat from scientific thinking, from which we are now just recovering.

The lesson seems to be that scientific work can inform but never replace the judgement of the policy-makers, practitioners, and parents. We can do much better than we have done in making scientific information available, but if the contribution of research is to be sustained, we must be careful not to oversell it. Perhaps the best safeguard against overselling is strong peer review. Scientists are trained skeptics and a healthy dose of skepticism keeps the enterprise healthy, spurring new investigations while constraining unwarranted generalizations.

Conclusions

1. Scientific credibility in educational research is no different from scientific credibility in health research. Four years on an NIH peer-review committee convinced me that top researchers in pediatrics, linguistics, developmental psychology, statistics, psychiatry, and education use essentially similar norms in evaluating the credibility of scientific claims and new research proposals.

2. In the final analysis, it is the peer review process within the scientific community that tells society when a claim is backed by science. If we want to improve scientific inquiry in education we must improve peer review. Peer reviewers in NIH are remarkably committed to principles of objectivity -- to incredibly careful scrutiny of alternative explanations and evidence. We should set the same standard for peer review in education.

3. Scientific inquiry in education, however, is not cheap. An experiment that assigns schools to whole-school reform programs is a large-scale enterprise. The fraction of educational spending that goes to research is, however, tiny as compared to the fraction of the health care budget that goes to health research. It is hard to imagine how the educational research enterprise, including high-level peer review, can improve without more funding.

4. Scientific research in education takes many forms: large-scale surveys, small-scale qualitative inquiry, and experimental or non-experimental evaluations of new programs. However, in my view, our research agenda has been out of balance in recent decades. *Making valid causal inferences about the impacts of our interventions* is, in my view, the key challenge facing us now. Lots of good work using surveys and qualitative inquiry can help us identify unsolved problems -- that is, targets of intervention, and also promising new ideas about practice. At the end of the day, however, we must judge our research enterprise by its track record in sorting out claims about the impact of educational interventions on student learning.

5. Randomized experiments are powerful tools for evaluating causal claims. We ought to find ways of doing more experiments.

6. However, well-designed non-experimental studies can also be effective and are sometimes the only way to assess impact. A recent conference called by Secretary Paige considered opportunities for learning "what works" by exploiting the availability of annual testing data on students. Researchers at the Consortium for School Research in Chicago have led the way in this regard. They have shown how annual testing data on multiple cohorts of students can be used to assess the impact of a new policy that ends social promotion (Consortium on Chicago School Research, 1999). This kind of work requires considerable research skill but can be extremely cost effective.

7. Let's keep our aims for scientific contributions to education realistic. If we oversell

what science can do, we set the stage for cynicism and a long-term decline in support for research.

8. Finally, lots of people think they know how to reform education. We've all been in school and we think we know what works. Teaching, however, is a demanding and complex activity, and organizing schools to support good instruction is equally challenging. Though educational research lacks the specialized language and complex equipment used in medical research, disciplined inquiry guided by critical scrutiny of truth claims is no less important. I am delighted and thankful to participate in a meeting such as this where these principles are taken seriously.

References

Consortium on Chicago School Research. (1999). *Ending social promotion: Results from the first two years* (M. Roderick, A. S. Bryk, B. A. Jacob, J. Q. Easton, & E. Allensworth, Trans.). Chicago: University of Chicago.

Cook, T. D. (2001). Considering the major arguments against random assignment: An analysis of the intellectual culture surrounding evaluation in American schools of education. In R. Boruch & F. Mosterller (Eds.), *Education, evaluation and randomized trials.* Brookings.

Cook, T. D., Habib, F. N., Phillips, M., Settersten, R. A., Shagle, S. C., & Degirmencioglu, S. M. (1999a). Comer's school development program in Prince George's County, Maryland: A theory-based evaluation. *American Educational Research Journal, 36*(3), 543-598.

Cook, T. D., Hunt, H. D., & Murphy, R. F. (1999b). *Comer's school development program in Chicago: A theory-based evaluation.*, Northwestern University.

Meier, P. (1972). The biggest public health experiment ever: The 1954 field trial of the Salk poliomyelitis vaccine. In J. M. Tanur, F. Mosteller, W. H. Kruskal, R. F. Link, R. S. Pieters & G. R. Rising (Eds.), *Statistics: A guide to the unknown* (pp. 2-13). San Francisco: Holden-Day, Inc.

Raudenbush, S. W. (1982). *Two scientific revolutions that failed: What can we learn from them about how social science can contribute to practice?*. Unpublished paper, Harvard Graduate School of Education, Cambridge, MA.

Slavin, R., & Madden, N. (In press). *One million children: Success for all.* Thousand Oaks, CA: Corwin.

Snow, C. E., Burns, M. S., & Griffin, P. (1998). *Preventing reading difficulties in young children* (C. E. Snow, M. S. Burns, & P. Griffin, Eds.). Washington D.C.: National Academy Press.

Thorndike, E. (1910). The contribution of psychology to education. *Journal of Educational Psychology, 1*(1), 8.

Whiteley, B. J., Weinshenker, M., & Seelig, S. E. (2002). *The AERA Research Grants Program: Key findings of selected tudies* (A report to the AERA Grants Board). Chicago, Illinois.