Evaluating What Works in Education: Better Methods and Wiser Consumers.

Phoebe H. Cottingham, Commissioner
Education Evaluation and Regional Assistance
Institute of Education Sciences, U.S. Department of Education

Why the call for better evidence about education policies?
What can be done to produce better evidence?

The overall purpose here is to look at what governments are doing to systematically evaluate what works in education. I want to summarize what and why IES – the new Institute of Education Sciences charged with bringing more scientific rigor to education evaluation – is doing in this area.

The call for evidence about what works

Much of the interest, really demand, for more objective, scientific evidence about the effects of policies comes from our belief that public resources should be *efficiently* dedicated to whatever the purposes established through the policy process.

Accountability – marking achievement of purposes – relies on *evidence* of achievement.

Then the challenging task is making good choices about what actions will move the outcome indicators. Evidence about the outcomes likely from alternative actions is science-based if scientific principles are followed in setting up the accountability structure and implementing it – measures must adequately capture the outcomes sought and be applied consistently.

Is observational science enough to determine effectiveness?

Observational science – measuring outcomes across units (e.g. schools, grades, students) and through time – is the starting point and essential under most accountability systems. Observations tell us where we are and what the trend looks like.

Trouble begins when we rely only on observational science to diagnose and treat the system – for example, assuming some aspect of a teacher's training is what accounts for apparent success or failure with children in the classroom.

Many causal relationships are suggested in what can be observed. It is easy to presume a causal mechanism is at work. One may conclude from observing one group of high-performing students that something in the pedagogy or classroom setting is critical. Applications of that feature for another group of students may not produce the same outcomes. The real causative agent could one or more unobservable factors. This is frustrating, but reality.

It has been shown – more on this topic later today – that one will get different answers when one directly tests the effect of a practice or policy using experimental design to assure that all the unobservable and observable are equally – randomly – distributed across those experiencing the new practice or program and those who are placed in a control group.

A short history of how experimental methods became essential tools for evidence-based policy analysis.

Evidence-based policy analysis began in earnest in the 1960s with ambitious efforts to open educational opportunities and reduce poverty. The tools of social science were new and untried. One successful attempt was the Negative Income Tax Experiment. In education early attempts focused on Head Start and some other preschool program models, but these large-scale studies were weak in design and execution.

By the 1990s, the tools developed after two decades of experimentation in social welfare and job training – to a lesser extent, early childhood interventions – were tried in a few education experiments, such as Upward Bound (a federally-funded program that bridges high schoolers from urban poverty schools to college), Career Academies (special drop-out prevention program), and a national assessment of school drop-out prevention programs.

These efforts were similar to the Department of Labor's youth unemployment experiments. They relied on the older, voluntary student population, avoiding the issues of how to utilize experimental designs at the classroom or teacher level.

Thus large-scale and small-scale trials of education policies and practices in the K-12 years were still a rarity with one exception – the class-size experiment in Tennessee was remarkable as a study created by a state legislature to resolve fights over school financing. For the most part, both government and private foundations avoided requiring experimental designs for the many highly visible school reform initiatives that came and "went."

The only exception was an independent evaluation by Tom Cook of the effect of the James Comer whole school reform model in selected school systems. Cook succeeded in carrying out a limited experiment – limited to several school districts – and did not find evidence of distinctive effects due to weak implementation.

In most cases where significant amounts of money were being invested, funding went directly to school systems to take on a different management approach, or to intermediary organizations who were supposed to influence school performance or bring in community services that might improve child learning or support their parents. Corporate philanthropy blossomed, but these efforts tended to have no empirical testing. Some

were quite large-scale – such as the Annenberg School project – but were not designed or initiated to seriously test if a school reform idea "worked."
Un-noticed to funders of education research was the maturation of an independent social science research sector located outside the academy. This sector expanded as the federal government funded large-scale experiments in social welfare and job training. The research firms and organizations that formed the nucleus drew their staff from those trained in economics, sociology, and psychology (especially those trained in statistics and empirical research methods).

Curiously, most of the new breed of science-based policy research "houses", while based outside of academy, hired researchers from disciplines outside of education, rather than scholars involved in the traditional professional schools dedicated to teacher and school administrator training.

This pattern is true for not only the empirical evaluation work but also for the generation of new policy ideas. The more innovative ideas about school improvement and learning come from those outside the professional education system – from political scientists, economists, psychologist, and others. Typically these individuals are highly skilled in observational assessment or in framing educational problems as systemic issues. Some are engaged in testing new cognitive science applications in education.

At NICHD – separate from the U.S. Department of Education -- Reid Lyon, a specialist in learning behaviors, led others in conducting small-scale trials on reading, illuminating the role played by early language acquisition. These findings formed the basis for Reading First, a federally funded program to encourage schools to utilize the new science and revise their teaching methods.

Others favor changing incentive systems. This line of evidence-based policy research is focused on competition in education, especially testing policies aimed to expand school choice. Others focus on particular pieces of the picture – teacher unions, pay systems, and preparation through state accreditation agencies, including testing alternative ways to do teacher training and institute performance systems. The creation of charter schools – public schools that are not under the local school district authority but potentially competitive with existing schools are another focus for evaluation. Overall with so many new ideas for school improvement, the great need is to assess these ideas through strong evaluations.

New initiatives at IES

At IES's National Center for Education Evaluation and Regional Assistance, there are three main lines of work:

- *Production of new evidence* by sponsoring randomized controlled trials on the most critical questions confronting educators. IES is deploying federal resources

for RCTs in priority areas identified by Congress, the Administration, and other stakeholders in our education system.

- *Establish the baseline of what we know* from existing evaluations. Most efforts to review and synthesize evidence are sporadic and highly individualized; IES is developing the What Works Clearinghouse to apply scientific standards consistently and continuously across studies on priority topics to reveal where the evidence is strong and where it is weak.

- *Develop an informed user constituency* that can explain and utilize the results from policy research.

Production of new evidence

A critical objective of the Institute of Education Sciences is to produce much better evidence on the effectiveness of policy options before the proposed practices are instituted into universal policy or practice.

The National Center for Education Evaluation started *ten* new rigorous evaluations in 2003, using the most scientifically credible methods. It is likely that 6 or 7 additional experiments could be announced in the next year.

These are the new studies going into the field:

1. **Education Technology** – Field-testing of 17 computer-based learning systems selected through a competitive process begins this fall in schools. Students are assigned to classrooms and classrooms/teachers are assigned to using the technology packages or following regular practice.
2. **Power4Kids** -- a test of brand-name reading programs designed to help the lowest functioning readers in the primary grades – this unique project will be presented in one of the meeting session.
3. **Early Reading First** – evaluation of federal support for early literacy curriculums in preschool programs begins in the fall '04.
4. **Reading First** – evaluation of major federal support to schools to introduce stronger reading instructions in the primary grades.
5. **English Language Learners** – testing enhanced versions of structured English immersion and transitional bilingual approaches in elementary grades.
6. **After-school Academic Instruction** – two developers are creating new curriculums appropriate for after-school programs, with a pilot test next year, and full scale evaluation the year after.
7. **DC School Choice Program** – this federally funded scholarship program will be rigorously evaluated.
8. **Charter Schools** – the evaluation will use entrance lotteries to study the effects of studying in a charter school.
9. **Teacher Preparation** – an experimental evaluation of classroom effects of traditional teacher preparation programs and less traditional preparation systems.

10. **Teacher Professional Development** – experimental test of more intensive methods to improve performance.

The **Even Start** trial – adding early childhood education curriculums to the Even Start program – began in 2002. NCEE is also winding up this year the second-year findings for the **21st Century Community Learning Centers** evaluation, following the release last year of the first-year findings.

Establish a systematic review center to bring forth the strongest evidence on interventions.

The Institute of Education Sciences started a new entity, the What Works Clearinghouse (WWC) in September 2002. The purpose is to provide a "central, trusted source of evidence about what works in education policy, practice, programs, and interventions." A joint venture formed by American Institutes of Research and The Campbell Collaboration with three subcontractors was awarded a five-year contract in 2002 to take responsibility for creating and maintaining WWC.

WWC set forth stringent standards for its work. The WWC Technical Advisory Group, chaired by Professor Larry Hedges, with thirteen other specialists in evaluation methodology and synthesis, advises on standards for selecting studies and assessing them.

WWC recently expanded the review system to meet the overwhelming demand for evidence about particular interventions being considered by educators to improve student learning. To accommodate these interests, WWC searches through all extant studies and identified first those that follow RCT designs. These studies are carefully reviewed, with reports on the studies and interventions. All information will be located on the WWC website, opening May 12th. Users will be able to search for interventions by name, to learn if WWC has found any strong evidence in support of effectiveness. As new studies are sent in, WWC plans to add them to the review process, and issue new reports monthly.

WWC rates studies primarily on study design – sorting them into three categories, (1) evidence level 1, or "meets WWC standards", meaning the study used a randomized controlled trial (RCT) design; (2) evidence level 2, or "meets WWC standards with reservations", meaning some type of high-quality comparison group design was used; and (3) no evidence, or "fails to meet WWC standards".

What has been learned so far in this endeavor? First, that there is tremendous demand for clear, objective assessment of studies that claim to have produced evidence about an intervention's effectiveness. Second, there is also strong interest on the part of product developers and evaluators to have their particular studies assessed. Third, those responsible for carrying out government policies and having to decide what to fund as evidence-based services or products, expect to use WWC findings across a very large

spectrum of interventions. Since many interventions have no credible study there will likely be consternation about the lack of evidence.

Studies that most easily pass the WWC standards will be RCTs but the answers produced by these studies may not seem sufficiently impressive to some advocates. Most studies report findings as effect sizes or changes in some outcome variable. Policy makers then have to decide if that amount of change is worth the cost of the intervention. The amount of change may seem trivial unless interpreted by projections of possible longer-term effects to weigh against costs.

Develop an informed user constituency

Government support for education services is under much closer scrutiny, as passage of No Child Left Behind made remarkably clear. Therefore, evaluation and knowledge utilization work at IES is aimed to first satisfy the needs of decision makers in education, who must have an evidence base for what they do.

Practitioners also need to become savvier about how to use monitoring systems and interpret results from RCTs. More work is needed to help practitioners understand how to interpret study findings and develop stronger applications of the findings into accountability systems. The Practitioner's Guide prepared by Jon Baron gives one a good start. The systematic review of replication studies – studies that use correlational science to approximate experimental studies – is another window into understanding why one needs an experiment when knowing what works is important.

Overall, broader understanding of the difference between correlational science and experimental science is needed. Practitioners must also have the flexibility to change practices, especially if rigorous evaluations make a case for change in one direction or another.