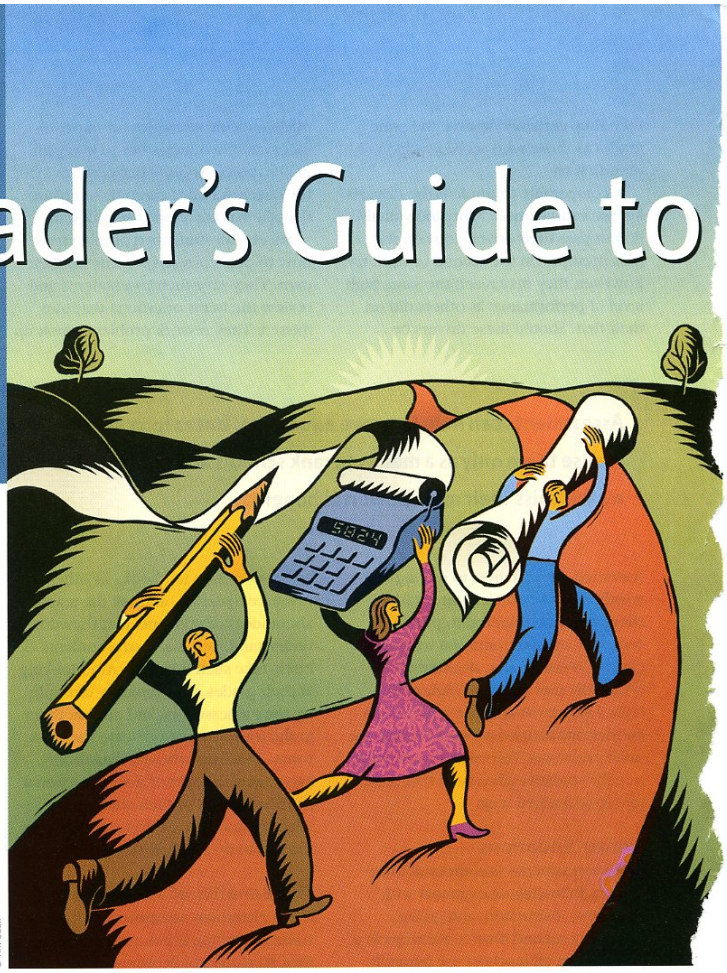
A Reader's Guide to

Learning how to assess the validity of education research is vital for creating effective, sustained reform.



© Tim Cook

Robert E. Slavin

In every successful, dynamic part of our economy, evidence is the force that drives change. In medicine, researchers continually develop medications and procedures, compare them with current drugs and practices, and if they produce greater benefits, disseminate them widely. In agriculture, researchers develop and test better seeds, equipment, and farming methods. In technology, in engineering, in field after field, progress comes from research and development. Physicians, farmers, consumers, and government officials base key decisions on the results of rigorous research.

In education reform, on the other hand, research has played a relatively minor role. Untested innovations appear, are widely embraced, and then disappear as their unrealistic claims fail to materialize. We then replace them with equally untested innovations diametrically opposed in philosophy, in endless swings of the reform pendulum. Far more testing goes into our students' hair gel and acne cream than into most of the curriculums or instructional methods teachers use. Yet which of these is more important to our students' future?

Evidence-Based Reform

At long last, education reform may be entering an era of well-researched programs and practices (Slavin, 2002). The U.S. government is now interested in the research base for programs that schools adopt. The Comprehensive School Reform Demonstration legislation of 1997 gives grants to schools to adopt "proven, comprehensive" reform designs. Ideally, "proven" means that programs have been evaluated in "scientifically based research," which is defined as "rigorous, systematic, and objective procedures to obtain valid knowledge" (U.S. Department of Education, 1998). The emphasis is on evaluations that use experimental or quasi-experimental designs, preferably with random assignment. The Bush administration's No Child Left Behind Act mentions "scientifically based research" 110 times in references to Reading First programs for grades K–3, Early Reading First for preK, Title I school improvement programs, and many more. In each case, schools, districts, and states must justify the programs that they expect to implement under federal funding.

Judging the Validity of Education Research

The new policies that base education funding and practice on scientifically based, rigorous research have important consequences for educators. Research matters. Educators have long given lip service to research as a guide to prac-

Scientifically Based Research

tice. But increasingly, they are being asked to justify their choices of programs and practices using the findings of rigorous, experimental research.

Why is one study valid whereas another is not? There are many valid forms of research conducted for many reasons, but for evaluating the achievement outcomes of education programs, judging research quality is relatively straightforward. Valid research for this purpose uses meaningful measures of achievement to compare several schools that used a given program with several carefully matched control schools that did not. It's that simple.

Control Groups

A hallmark of valid, scientifically based research on education programs is the use of control groups. In a good study, researchers compare several schools using a given program with several schools not using the program but sharing similar demographics and prior performance, preferably in the same school district. Having at least five schools in each group is desirable; circumstances unique to a given school can bias studies with just one or two schools in each group.

A control group provides an estimate of what students in the experimental program would have achieved if they had been left alone. That's why the control schools must be as similar as possible to the program schools at the outset.

Randomized and Matched Experiments

The most convincing form of a control group comparison is a randomized experiment in which students, teachers, or schools are assigned by chance to a group. For example, the principals and staffs at ten schools might express interest in using a given program. The schools might be paired up and then assigned by a coin flip to the experimental or control group.

Randomized experiments are very rare in education, but they can be very influential. Perhaps the best known example in recent years is the Tennessee class size study (Achilles, Finn, & Bain, 1997/1998) in which researchers assigned students at random to small classes (15 students), regular classes (20–25 students), or regular classes with an aide. The famous Perry Preschool Program (Berrueta-Clement, Schweinhart, Barnett, Epstein, & Weikart, 1984) assigned four-year-olds at random to attend an enriched preschool program or to stay at home. Two recent studies of James Comer's School Development Project randomly assigned schools to use the School Development Project or keep using their current program (Cook et al., 1999; Cook, Murphy, & Hunt, 2000). In each of these studies, random assignment made it very likely that the experimental and control groups were identical at the outset, so any differences at the end were sure to have resulted from the program.

Matched studies are far more common than randomized ones. In a matched program evaluation, researchers compare students in a given program with those in a control group that is similar in prior achievement, poverty level, demographics, and so on. Matched studies can be valid if the experimental and control groups are very similar. Often, researchers use statistical methods to "control for" pretest differences between experimental and control groups. This can work if the differences are small, but if there are large differences at pretest, statistical controls or use of test-gain scores (calculated by subtracting pretest scores from posttest scores) are generally not adequate.

The potential problem with even the best matched studies is the possibility that the schools that chose a given program have (unmeasured) characteristics that are different from those that did not choose it. For example, imagine that a researcher asked 10 schools to implement a new program. Five enthusiastically take it on and five refuse. Using the refusal group as a control group, even if it is similar in other ways, can introduce something called selection bias. In this example, selection bias would work in favor of finding a positive treatment effect because the volunteer schools are more likely to have enthusiastic, energetic teachers willing to try new methods than are the control schools. In other cases, however, the most

desperate or dysfunctional schools may have chosen or been assigned to a given program, giving an advantage to the control schools.

Is Random Assignment Essential?

Random assignment to experimental and control groups is the gold standard of research. It virtually eliminates selection bias because students, classes, or schools were assigned to treatments not by their own choice but by the flip of a coin or another random process.

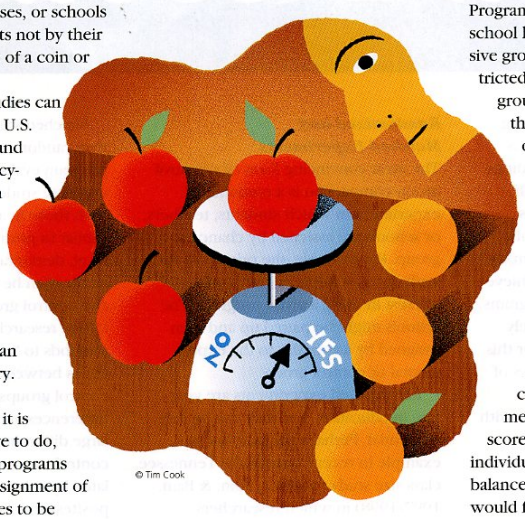
Because randomized studies can rule out selection bias, the U.S. Department of Education and many researchers and policymakers have recently been arguing for a substantial increase in the use of randomized designs in evaluations of education programs. Already, more randomized studies are under way in education than at any other point in history.

The only problem with random assignment is that it is very difficult and expensive to do, especially for schoolwide programs that necessitate random assignment of whole schools. No one likes to be assigned at random, so such studies often have to provide substantial incentives to get educators to participate. Still, such studies are possible; we have such a study under way to evaluate our Success for All comprehensive reform model, and, as noted earlier, Comer's School Development Program has been evaluated in two randomized studies.

At present, with the movement toward greater use of randomized experiments in education in its infancy, educators evaluating the research base for various programs must look carefully at well-matched experiments, valuing those that try to minimize bias by using closely matched experimental and control groups, having adequate numbers of schools, avoiding comparing volunteers with nonvolunteers, and so on.

Statistical and Educational Significance and Sample Size

Reports of education experiments always indicate whether a statistically significant difference exists between the achievement of students in the experimental group and those in the control group, usually controlling for pretests and other factors. A usual criterion is "$p < 0.05$," which means that the probability is less than 5 percent that an observed difference might have happened by chance.



© Tim Cook

The proportion of students within a program getting "significantly higher" scores than those in a control group is important, but it may not be important enough. In a large study, a small difference could be significant. A typical measure of the size of a program effect is "effect size," the experimental-control difference divided by the control group's standard deviation (a measure of the dispersion of scores). In education experiments, an effect size of +0.20 (20 percent of a standard deviation) is often considered a minimum for significance; effect sizes above +0.50 would be considered very strong.

But student groupings can have a profound impact on student outcomes. Often, an experiment will compare one school using Program *X* with one matched control school. If 500 students are in each school, this is a very large experiment. Yet the difference between the Program *X* school and the control school could be due to any number of factors that have nothing to do with Program *X*. Perhaps the Program *X* school has a better principal or a cohesive group of teachers or has been redistricted to include a higher-performing group of students. Perhaps one of the schools experienced a disaster of some sort—in an early study of our Success for All program, Hurricane Hugo blew the roof off of the Success for All school but did not affect the one control school.

Because of the possibility that something unusual that applies to an entire school could affect scores for all the students in that school, statisticians insist on using the *school's* means, not individual student scores, in their analyses. In this way, individual school factors are likely to balance out. Statistical requirements would force a researcher to have at least 20–25 *schools* in each condition. Very few education experiments are this large, however, so the vast majority of experiments analyze at the student level.

Readers of research must apply a reasonable approach to this problem. We should view studies that observe a single school or class for each condition with great caution. However, a study with as many as five program schools and five control schools probably has enough schools to ensure that a single unusual school will not skew the results. Such a study would still use individual scores, not school means, but it would be far preferable to a comparison between only two schools.

A single study involving a small number of schools or classes may not be conclusive in itself, but many such studies, prefer-

ably done by many researchers in a variety of locations, can add confidence that a program's effects are valid. In fact, experimental research in education usually develops in this way. Rather than evaluate one large, definitive study, researchers must usually look at many small studies that may be flawed in various (unbiased) ways. But if these studies tend to find consistent effects, the entire set of studies may produce a meaningful conclusion.

Research to Avoid

All too often, program developers or advocates cite evidence that is of little value or that is downright misleading. A rogue's gallery of such research follows.

Cherry Picking

Frequently, program developers or marketers report on a single school or a small set of schools that made remarkable gains in a given year. Open any education magazine and you'll see an ad like this: "Twelfth Street Elementary went from the 20th percentile to the 60th in only one year!" Such claims have no more validity than advertisements for weight loss programs that tell the story of one person who lost 200 pounds (forgetting to mention the hundreds who did not lose weight on the diet). This kind of "cherry picking" is easy to do in a program that serves many schools; there are always individual schools that make large gains in a given year, and the marketer can pick them after the fact just by looking down a column of numbers to find a big gainer. (Critics of the program can use the same technique to find a big loser.) Such reports are pure puffery, not to be confused with science.

Bottom Fishing

A variant of cherry picking is "bottom fishing," using an after-the-fact comparison in which an evaluator compares schools using a given program with matched "similar schools" known to have made poor gains in a given year. Researchers can legitimately compare gains made in program schools and gains made in the entire district or state because the large comparison group makes "bottom fishing" impossible. However, readers should interpret with caution after-the-fact studies purporting to compare groups selected by the evaluator.

Pre–Post Studies

Another common but misleading design is the pre-post comparison, lacking a control group. Typically, the designer cites standardized test data, with the rationale that the expected year-to-year gain in percentiles, normal curve equivalents, or percent passing is zero, so any school that gained more than zero has made good progress.

The problem with this logic is that many states and districts make substantial gains in a given year, so the program schools may be doing no better than other schools. In particular, states usually make rapid gains in the years after they adopt a new test. At a minimum, studies should compare gains made in program schools in a given district or state with the gains made in the entire district or state.

Random assignment to experimental and control groups is the gold standard of research.

Scientifically Based Versus Rigorously Evaluated

A key issue in the recent No Child Left Behind legislation is the distinction between programs that are "based on scientifically based research" and those that have been evaluated in valid scientific experiments. A program can be "based on scientifically based research" if it incorporates the findings of rigorous experimental research. For example, reading programs are eligible for funding under the federal Reading First initiative if states determine that they incorporate a focus on five elements of effective reading instruction: phonemic awareness, phonics, fluency, vocabulary, and comprehension. The National Reading Panel (1999) identified these elements as having been established in rigorous research, especially in randomized experiments. Yet there is a big difference between a program *based* on such elements and a program that has itself been compared with matched or randomly assigned control groups. We can easily imagine a reading program that would incorporate the five elements but whose training was so minimal that teachers did not implement these elements well, or whose materials were so boring that students were not motivated to study them.

The No Child Left Behind guidance (U.S. Department of Education, 2002) recognizes this distinction and notes a preference for programs that have been rigorously evaluated, but also recognizes that requiring such evaluations would screen out many new reading programs that have not been out long enough to have been evaluated, and so allows for their use. This approach may make sense from a pragmatic or political perspective, but from a research perspective, a program that is unevaluated is unevaluated, whether or not it is "based on" scientifically based research. A basis in scientifically based research makes a program promising, but not proven.

Research Reviews

In order to judge the research base for a given program, it is not necessary that every teacher, principal, or superintendent carry out his or her own review of the literature. Several reviews applying standards have summarized evidence on various programs.

For comprehensive school reform models, for example, the American Institutes for Research published a review of 24 programs (Herman, 1999). The Thomas Fordham Foundation (Traub, 1999) commissioned an evaluation of 10 popular comprehensive school reform

models. And Borman, Hewes, Rachuba, and Brown (2002) carried out a meta-analysis (or quantitative synthesis) of research on 29 comprehensive school reform models.

Research reviews facilitate the process of evaluating the evidence behind a broad range of programs, but it's still a good idea to look for a few published studies on a program to get a sense of the nature and quality of the evidence supporting a given model. Also, we should look at multiple reviews because researchers differ in their review criteria, conclusions, and recommendations. Adopting a program for a single subject, much less for an entire school, requires a great deal of time, money, and work—and can have a profound impact on a school for a long time. Taking time to look at the research evidence with some care before making such an important decision is well worth the effort. Accepting the developer's word for a program's research base is not a responsible strategy.

How Evidence-Based Reform Will Transform Our Schools

The movement to ask schools to adopt programs that have been rigorously researched could have a profound impact on the practice of education and on the outcomes of education for students. If this movement prevails, educators will increasingly be able to choose from among a variety of models known to be effective if well implemented, rather than reinventing (or misinventing) the wheel in every school. There will never be a guarantee that a given program will work in a given school, just as no physician can guarantee that a given treatment will work in every case. A focus on rigorously evaluated programs, however, can at least give school staffs confidence that their efforts to implement a new program will pay off in higher student achievement.

In an environment of evidence-based reform, developers and researchers will continually work to create new models and improve existing ones. Today's substantial improvement will soon be replaced by something even more effective. Rigorous evaluations will be common, both to replicate evaluations of various models and to discover the conditions necessary to make programs work. Reform organizations will build capacity to serve thousands of schools. Education leaders will become increasingly sophisticated in judging the adequacy of research, and, as a result, the quality and usefulness of research will grow. In programs such as Title I, government support will focus on helping schools adopt proven programs, and schools making little progress toward state goals may be required to choose from among a set of proven programs.

A control group provides an estimate of what students in the experimental program would have achieved if they had been left alone.

Evidence-based reform could finally bring education to the point reached early in the 20th century by medicine, agriculture, and technology, fields in which evidence is the lifeblood of progress. No Child Left Behind, Reading First, Comprehensive School Reform, and related initiatives have created the possibility that evidence-based reform can be sustained and can become fundamental to the practice of education. Informed education leaders can contribute to this effort. It is ironic that the field of education has embraced ideology rather than knowledge in its own reform process. Evidence-based reform honors the best traditions of our profession and promises to transform schooling for all students. ■

Author's Note: This article was written under funding from the Office of Educational Research and Improvement, U.S. Department of Education. Any opinions expressed are those of the author and do not necessarily represent OERI positions or policies.

References

Achilles, C. M., Finn, J. D., & Bain, H. P. (1997/1998). Using class size to reduce the equity gap. *Educational Leadership, 55*(4), 40–43.

Berrueta-Clement, J. R., Schweinhart, L. J., Barnett, W. S., Epstein, A. S., & Weikart, D. P. (1984). *Changed lives*. Ypsilanti, MI: High/Scope.

Borman, G. D., Hewes, G. M., Rachuba, L. T., & Brown, S. (2002). *Comprehensive school reform and student achievement: A meta-analysis*. Submitted for publication. (Available from the author at gborman @education.wisc.edu)

Cook, T. D., Habib, F., Phillips, M., Settersten, R. A., Shagle, S., & Degirmencioglu, M. (1999). Comer's school development program in Prince George's County, Maryland: A theory-based evaluation. *American Educational Research Journal, 36*(3), 543–597.

Cook, T., Murphy, R. F., & Hunt, H. D. (2000). Comer's school development program in Chicago: A theory-based evaluation. *American Educational Research Journal, 37*(2), 543–597.

Herman, R. (1999). *An educators' guide to schoolwide reform*. Arlington, VA: Educational Research Service.

National Reading Panel. (1999). *Teaching children to read*. Washington, DC: U.S. Department of Education.

Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher, 31*(7), 15–21.

Traub, J. (1999). *Better by design? A consumer's guide to schoolwide reform*. Washington, DC: Thomas Fordham Foundation.

U.S. Department of Education. (1998). *Guidance on the comprehensive school reform demonstration program*. Washington, DC: Author.

U.S. Department of Education. (2002). *Draft guidance on the comprehensive school reform program* (June 14, 2002 update). Washington, DC: Author.

Robert E. Slavin is the codirector of the Center for Research on the Education of Students Placed At Risk at Johns Hopkins University and the chairman of the Success for All Foundation, 200 W. Towsontown Blvd., Baltimore, MD 21204; rslavin@successforall.net.