The Principle That Social Interventions ought to be evaluated has a long pedigree. Eager readers of the Muqaddimah know that Ibn Khaldun considered competing explanations for the success of Arab regimes in the 13th century. In the 19th century, Florence Nightingale reproved the English Parliament for failing to weigh seriously the consequences of making changes to social programs, charging that “without an inquiry after results, past or present, it is all experiment, see-saw, doctrinaire, a shuttlecock between two battledores.” A 21st-century Nightingale could testify at a congressional hearing with only slight revisions. After a 1995 review of evidence on the performance of federal employment and training programs, the General Accounting Office concluded that “most federal agencies do not know if their programs are working.”

Nevertheless, the 20th century brought great progress in the theory and methods of evaluation as well as more understanding of its necessity. The most promising innovation was the randomized field trial. Such experiments randomly “treat” individuals and even whole institutions, such as hospitals or schools, with different interventions in order to learn which work better. The random allocation ensures that the two groups being compared are not different in ways that would influence their response to a particular treatment. For instance, suppose you wanted to test the quality of a new reading curriculum. You might select some students to be taught using the new curriculum and compare their progress with that of students who stayed with the old curriculum. But the students chosen to receive the new curriculum might differ in a way that influences their academic

by ROBERT BORUCH

For years, researchers have used randomized field trials to evaluate the effectiveness of medical procedures.

Educators are finally starting to catch on
progress—they might, for instance, have had better reading teachers in the past, be more motivated, or have access to more educational resources at home. Randomly allocating students or multiple classrooms or multiple schools to the new reading program and to control conditions or to an alternative program eliminates the possibility that the two groups will differ in a systematic way and thus compromise the results.

Randomized field trials are a sturdy method of generating defensible evidence about the relative effectiveness of various interventions; nonrandomized trials do so, at times, but unpredictably. For instance, in the Salk polio vaccine studies of the 1950s, randomized trials that were mounted in some states produced estimates of the vaccine's effect on polio that were appreciably greater than estimates from a parallel series of nonrandomized trials. The factors that might have led to this difference are still not well understood. After the Salk trials, controversy erupted over the use of oxygen-enrichment therapy for premature infants. Early nonrandomized studies suggested that the infant death rate was reduced significantly by the oxygen-enrichment therapy. Subsequent randomized clinical trials helped to reveal that enriched oxygen environments for premature infants caused blindness and did not decrease infant mortality.

Yet this powerful technique of discovering what works has been slow to come to the field of education. Consider the widespread adoption of "whole school reform" models such as "Success for All," "Accelerated Schools," and "Expeditionary Learning." The promise of whole-school reform was to install standardized, high-performing, research-tested curricula in school after school, instead of forcing each school and school district to be its own curriculum developer. Much energy and expense has been devoted to implementing these prepackaged school-reform programs in mainly low-income school districts. Nevertheless, in 1999 the American Institutes of Research reviewed all the studies done on 24 reform models and found that 5 of the programs had no evidence beyond anecdote and personal testimonials to support their claims of raising achievement. The other 19 programs were subject to about 116 independent studies of their effectiveness. As far as can be determined from appendices in the report, only one of these studies, involving a test of a specialized instructional strategy in the context of the Paideia program, depended on a randomized trial.

**Inventive Research**

Before chastising education reformers, however, it is crucial to recognize how young an innovation the randomized field trial truly is. Even in fields where the use of randomized trials is now standard practice, such as medical research, their adoption was fairly recent. The earliest randomized trials, to test the effects of a diphtheria serum, were undertaken by the Danes in 1900. But the American and European research communities didn't regard randomized field trials as essential until the 1950s, when Jonas Salk discovered the polio vaccine and Sen. Estes Kefauver held hearings on the testing of thalidomide and other drugs. Now all pharmaceuticals must be subjected to rigorous clinical trials before being made available to the public. Research on industrial materials and chemical processes now depends in some measure on such trials as well.

Not just education but the social sciences in general lag behind the hard sciences in the use of randomized trials. Started in 1993, the international Cochrane Collaboration's library of existing randomized and possibly randomized trials on the effects of health-care interventions contains 250,000 entries. By contrast, a sister effort, the international Campbell Collaboration's Social, Psychological, Educational, and Criminological Trials Registry, begun in 1999, includes about 10,000 randomized and possibly randomized trials. More randomized trials will doubtlessly be unearthed as Campbell Collaboration researchers, mostly volunteers, continue to comb through the professional literature. But even if the number of trials were to increase five-fold, the social and behavioral sciences would remain relatively bereft of precise experimentation.

Two fundamental problems plague the use of randomized
trials in the social sciences. First, randomized trials can be an enormous undertaking; testing the effects of a job-training program or a new work incentive in welfare is not the same as giving one set of patients a new drug and another set a placebo. What measurements to use (future earnings? children’s health and welfare? happiness?) and how to isolate the effects of a given program versus all the other influences the world presents remain vexing questions in the design of randomized trials. In education, any number of factors can influence the outcome of a randomized trial. A new curriculum might be highly effective, but teachers may not be trained to use it properly. In order to obtain reliable results, policies must be implemented as they were designed to be. Medical researchers face similar issues, but not necessarily to the same degree as social scientists. The second problem is funding. Research on pharmaceuticals is funded by private firms with one goal in mind: bringing profitable drugs to market. They can pass their research costs onto consumers. Social scientists, by contrast, are most often testing the effects of programs funded by government or by philanthropic organizations.

Not until the 1960s, when Congress mandated that the newly created Head Start program be evaluated, was high-quality research performed in education. During this period, the Ypsilanti (Michigan) Preschool Demonstration Project was distinctive in using a randomized trial to estimate the High/Scope program’s effect on children’s achievement. But few other preschool programs were evaluated using randomized trials in the ‘60s; the Cochrane Collaboration database lists only six trials. The record hasn’t been much better in the decades since. A hand search of every article in every issue of the American Educational Research Journal since its inception revealed that of about 1,200 articles, only 35 concern randomized field trials in math or science education (see Figure 1). Moreover, there was no obvious increase in their number during the 1964 to 1998 period. This is very embarrassing; the education community is not doing much to generate solid evidence on what interventions work.

“Success for All” involved a randomized field trial.

Turning the Tide

There are reasons to be optimistic, though. In the current data-driven reform environment, more and more education policymakers are asking for reliable evidence of the effectiveness of new and existing policies and programs. At the same time, researchers have shown that high-quality randomized trials can be mounted in the education field. Some of the best trials lie at the intersection of education, social services, and juvenile-justice systems. Anthony Petrosino’s systematic reviews of randomized trials on the “Scared Straight” program, which involves prisoners lecturing at-risk youth about the consequences of crime, showed that its effects are at worst negative, at best negligible. Randomized trials on the D.A.R.E. program, which encourages students to stay away from drugs, showed the program was worthless in its mature multimillion-dollar form, all the while enjoying mostly uncritical support from police, teachers, and journalists. By 2001, D.A.R.E. sponsors and staffs had decided to modify the program in light of the findings.

Perhaps the most promising sign is that two of today’s liveliest debates—over the merits of class-size reduction and school choice—are informed by results from well-run randomized trials. Tennessee’s STAR experiment, which reduced class sizes in the early grades, established that a substantial reduction in class size yields significant gains in achievement. Independent reanalyses of the data by Frederick Mosteller, Alan Krueger, and others helped to verify the study’s original conclusions. These findings were largely ignored for five or more years after the results were published in the American Education Research Journal and assorted trade journals or reports. But by the late 1990s, at least a dozen state governors had built their education reform platforms around the idea, and it had become a federal priority under the

www.educationnext.org
Clinton administration. The STAR experiment shows how a single randomized field trial can begin to clarify the effect of a particular intervention against a backdrop of many nonrandomized trials. However, quasi-experimental evidence from Connecticut and other locations suggests that the Tennessee effects may not extend to all situations.

Findings from randomized trials of the privately funded voucher programs in New York City, Washington, D.C., and Dayton, Ohio, have drawn considerable press attention, and in spring 2002 they were mentioned in oral arguments before the U.S. Supreme Court. All these data have a way of elevating the discussion immeasurably. Once it is possible to estimate reliably the gains elicited by an intervention, the debate becomes more sophisticated and grounded than just Nightingale’s “shuttlecock between two battledores.”

The Bush administration’s current focus on reading was informed by a series of studies conducted under the aegis of the National Institutes of Child Health and Human Development. The work of Jack Fletcher, Barbara Foorman, and others depends on randomized trials, high-end nonrandomized trials, and scientifically conscientious looking around. The vicious debate over the merits of various methods of teaching reading was essentially settled when randomized trials showed that phonetic understanding was a necessary but not sufficient component of learning to read.

The U.S. Department of Education is increasingly recognizing the value of randomized trials. From 1995 to 1997 its Planning and Evaluation Service awarded 51 contracts to study the effects of federally sponsored programs, of which 5 involved randomized field trials on programs like Upward Bound, the Even Start Family Literacy Program, and the School Dropout Demonstration Assistance Program. These involve large commitments of personnel and funds over the course of several years. In 1996 the total amount awarded in contracts was about $18.6 million, of which $1.4 million was devoted to these randomized trials. Their amount may seem small, but the trials are multiyear. Thus the total commitment from 1991 to 1995 for the School Dropout Demonstration Assistance Program, involving a randomized trial in each of 16 sites, was $7.3 million. For Upward Bound trials, which involved 67 sites, the commitment exceeded $5.4 million from 1992 to 1996.

Many private foundations in the United States have no explicit policy on the evaluation of projects that they sponsor. As a consequence, relatively few foundations have funded randomized trials that generate evidence about whether the programs that they support actually work. Nonetheless, a few foundations seem admirable in subsidizing field tests that 1) a national, state, or local government is unable or unwilling to mount; and 2) build on government investments in randomized field trials to generate new knowledge. In the education arena, the Rockefeller Foundation’s support of randomized trials on the Minority Female Single Parent Program and of the program itself was remarkable. The William T. Grant Foundation has supported trials on mentoring, nurse visitation programs, and New Chance, a teen-parent demonstration program. The Smith Richardson Foundation has supported the previously mentioned randomized trials on voucher-based school choice as well as Krueger’s reanalysis of data from the Tennessee class-size trials.

However, foundations are still more willing to fund randomized experiments in medicine than in education. The Gates Foundation, for instance, funded by the multibillionaire founder of Microsoft, has admirably committed $50 million for a carefully designed randomized study of an AIDS prevention initiative in Africa. At the same time, none of the foundations’ investments in education, though similar in scale, use this powerful scientific technique. Given the foundation’s commitment to scientific research, the policy will undoubtedly evolve during the coming years.

The School as Subject

Most randomized field trials involve measuring the effects of an intervention on individual research participants. But in the education field especially, society is interested not only in individual effects but in the effects on schools and school districts of policy changes. A new kind of trial is randomly assigning whole organizations, including schools and school districts, to alternatives in order to test their effectiveness. This sounds impossible to some people, but there are more than a few good precedents. They anticipate the way in which trials on “whole school reform” programs, among other schoolwide or districtwide interventions, might be run.

In the health education arena, for example, researchers have contributed substantially to understanding when and how to run randomized trials using schools as the units of allocation and analysis to test schoolwide risk-reduction programs. Entire hospitals have been randomized in tests of ways to educate staff, including physicians, and to change the handling of certain illnesses. Housing projects and factories have been the units in trials on AIDS prevention initiatives that had education as a major component. The Rockefeller Foundation is sponsoring trials that involve randomizing housing projects to test the effects of integrated programs to enhance social, economic, and other kinds of capital in the projects.
The Campbell Collaboration's registry contains at least 50 entries on analogous trials in which conventional measures of academic achievement were the main outcome variable. Classrooms or schools have been used by Len Jason and his colleagues on tests of programs for transfer students, in Leonard Bickman's admirable but aborted trials on teacher-incentive programs in Tennessee, and in Sheppard Kellam's Baltimore studies on mental health and children's achievement. At least one sizable trial has been mounted to test a violence prevention program based on a set of elementary schools that were randomly assigned to the program and to a control condition. Thomas Cook and his colleagues set a precedent for testing a schoolwide reform program with his trials on James Comer's vaunted School Development Program. Two of these studies succeeded in randomly allocating eligible and willing schools to the Comer program and to control conditions and in producing unbiased estimates of the program's effect. The Chicago trial provided evidence that the program affected perceptions of school and academic climate and had a small effect on academic achievement over the period covered by the study. In Prince George's County, Maryland, no effects on beliefs or performance could be detected at the study's completion.

**Dissemination**

A rigorous, well-run randomized trial is useless if its results aren't made accessible to policymakers and the public. Two major efforts are under way to fill the gap between research and practice. In the health-care arena, the Cochrane Collaboration has created an international system to identify the best trials and to carry out an exquisitely conscientious review of them. Trustworthy studies are identified as such, while untrustworthy ones, despite their bright lights, are not. Since 1993 the Cochrane Collaboration has produced about 1,000 systematic reviews of studies on the effects of diverse health-care interventions. Hit the button on the web site, ask about your illness, and find out what the reviews say based on good trials. Roughly 2,000 people from 40 countries are involved in this effort. It is a good sign of people's interest in evidence about what works in this arena.

In the late 1990s, the Cochrane people debated whether to enlarge the group's scope. The conversations led to the creation of the international Campbell Collaboration, the aim of which is to produce systematic reviews of high-quality studies of the effects of new approaches to education, crime and justice, and social welfare. The Campbell Collaboration focuses first on randomized trials; its secondary focus is nonrandomized trials. This organization transcends disciplinary boundaries and geopolitical limits and promises enhanced attention to what works and what does not. Both collaborations have depended heavily on voluntary efforts in their development, a good sign of common interest in their aims. Moreover, collaborators hail from different disciplines. For education researchers, this makes it easier to learn about studies of crime interventions in which education outcomes were measured.

Randomized trials will only become more frequent in the education arena as time goes on. Today's reform environment demands hard evidence on which programs and policies are effective at raising student achievement. So far, randomized trials provide the best way of obtaining such evidence, and their techniques become more sophisticated by the day. The question is whether these results will be used to effect change or be held hostage to the ideological posturing that so often substitutes for evidence in the education world.

- Robert Boruch is a professor of education in the Graduate School of Education and of statistics in the Wharton School of Business at the University of Pennsylvania.