Boruch, R. (2007). Encouraging the flight of error: Ethical standards, evidence standards, and randomized trials. New Directions for Evaluation, 2007(113), 55-73.

3

Thomas Jefferson recognized the value of reason and scientific experimentation in the eighteenth century. This chapter extends the idea in contemporary ways to standards that may be used to judge the ethical propriety of randomized trials and the dependability of evidence on effects of social interventions.

Encouraging the Flight of Error: Ethical Standards, Evidence Standards, and Randomized Trials

Robert Boruch

In the eighteenth century, Thomas Jefferson wrote about freedom of religion, and possible religious intolerance. In his *Notes on the State of Virginia,* he argued against governmental preference for one faith over another. Jefferson (cited in Fabian, 1955) employed a scientific analog to make his case, saying first . . . , "The Newtonian principle of gravitation is now more firmly established . . . than it would be were the government to step in, and make it an article of necessary faith" (p. 136). And then, in speaking of Newton's scientific methods, he said, "Reason and experiment have been indulged, and error has fled before them" (p. 138).

Jefferson may not have gotten much political mileage out of this prose despite the speculation of the first sentence and the beauty of the second. Nonetheless, his sentences put him among statesmen who today would support evidence-based policy. They might also put him firmly in the ranks of conscientious contemporary trialists.

The Institute for Education Sciences (directly) and the Jerry Lee Foundation (indirectly) have provided recent support for my work on this topic. Neither organization is responsible for the views expressed here. I am indebted to Avital Darmon, a senior adviser to the Israel Academy of Sciences and Humanities, for retrieving the original Jefferson text here after I had given her the beautiful quote. I am indebted also to Nancy Brokaw, senior editor at Penn's Graduate School of Education, for good counsel on earlier versions of this chapter.

NEW DIRECTIONS FOR EVALUATION, no. 113, Spring 2007
Published online in Wiley InterScience (www.interscience.wiley.com) • DOI: 10.1002/ev.215

Jefferson's remarks are apropos to what follows. This chapter's first aim is to explain broad ethical standards for considering whether a particular kind of experiment, a randomized trial, can be construed as an option in studies that aim to estimate an intervention's effect. One of these standards includes the idea that methods other than randomized trials may produce equally defensible results. Partly as consequence, the second aim is to outline evidence on how the results of a randomized trial may differ from the results of other kinds of experiments.

The idea of randomized trials was not understood until the late nineteenth century, though Jefferson is likely to have appreciated it mainly because he abided by fairness, writ large, and had a proclivity for accurate numerical comparison. The third aim here is to update readers' understanding of contemporary standards of evidence and of related standards for reporting on randomized trials and quasi-experiments. It identifies entities that are advancing the state of the art in identifying and improving such trials.

Most of the illustrations here stem from evaluations undertaken in the United States. In the Jeffersonian spirit, however, this also depends on work in other democratic countries and by multinational entities. The opportunities for developing better evidence on what works are tantalizing. In what follows, I identify opportunities for understanding better the evidence in the context of federal and multinational initiatives.

Ethical Standards

Determining whether a randomized controlled trial can meet broad standards of social ethics in the United States and many other countries is critical. Determinations of this sort must, of course, precede assessments of a trial's feasibility. It has implications for a trial's desirability.

What ethical standards might be invoked, at the broadest level, to guide a determination of whether a trial is at least an acceptable option? The Federal Judicial Center (1981), the research arm of the superior courts in the United States, put a well-informed set of standards in declarative language. In what follows, standards are put into blunt and interrogatory form, based on Boruch (1997) and the references found there.

Simple questions and answers are put plainly first. Elaborated answers using contemporary trials as examples follow on the heels of difficult questions:

1. *Is the social problem serious?* If the answer is yes, then consider a randomized trial to evaluate purported solutions. Otherwise a trial is not worthwhile or ethical.
2. *Are purported solutions to the problem debatable?* If the answer is yes, then consider doing a randomized trial. If the answer is no, then adopt the purported solution.

3. *Will randomized trials yield more defensible (less equivocal and unbiased) results than alternative approaches to estimating effects?* If the answer is yes, consider mounting a randomized trial. If the answer is no, then rely on the alternative approach.
4. *Will the results be used?* If the answer is yes, then consider mounting a randomized trial. If the answer is no, be wary.
5. *Will human rights be protected?* If the answer is yes, then consider a randomized trial. If not, forget about the trial, or redesign the randomized trial so that rights are protected.

Consider next a concrete example to illustrate how the ethical questions might be addressed in a specific context. The example involves a randomized trial of conditional cash transfer programs such as Progresa in Mexico (Parker and Teruel, 2003, 2005) and elsewhere (Rawlings and Rubio, 2003; Rawlings, 2005). Progresa aimed to reduce the rate at which children drop out of school in order to work in the agricultural fields and in other jobs that contribute to family income.

The Problem and Its Seriousness. Children in poor villages in Mexico leave school early to work in the agricultural field for low wages. This income has been important, despite its low level, to augmenting family resources. The dropout rate in primary grades has been in the 40 to 70 percent range.

This loss in human intellectual capital is serious relative to national human capital interests. It is not necessarily a serious problem relative to local commercial interests in Mexico and in other countries. So there were disagreements about problem severity in Mexico, as there are elsewhere.

Debatable Solutions to the Problem. Purported solutions to this kind of problem are numerous. Arguments have been made for mandatory school enrollment and attendance programs. But how would one enforce the rules engendered by such programs in Mexico, much less the United States? National campaigns to inform parents about the national need to educate children are an option. But what effect do television or radio or Web-based campaigns have on behavior? Sanctioning poor parents or agrobusinesses (including small ones) or schools is an option, but the potential effects of sanctions are dubious.

Conditional cash transfer programs in Mexico, as in other countries, were also an option. Cash payments could be made to mothers of children, conditional on their children's school enrollment and school attendance, and other requirements. Here too program effects were debatable. Mexico chose this option and subjected it to empirical tests (in place of randomized trials) because the effects of the purported solution were debatable.

Evidence on the Effectiveness of Purported Solutions. Any purported solutions to the problem of enhancing human capital for Mexico had been debatable partly because direct evidence for their effects had usually been absent. Some experience could be drawn from other countries and was. For example, some welfare experiments in the United States have included

provisions for sanctioning welfare recipients for failing to ensure that their children attended school regularly partly because the evidence on effects of sanctions had not been clear. Sanctions did not work. Nonetheless, the experience of the United States and other democratic countries is not clearly relevant for Mexico, even when the experience was based on controlled trials in the 1990s.

More to the scientific point, and to people interested in dependable evidence, nobody could predict with any real certainty what would happen in the absence of the program at the local and regional levels in Mexico. Variation among families and villages, family behavior over time, temporal variation in jurisdiction resources, quality of records, and other stochastic influences complicate matters. Developing a defensible counterfactual, a fair comparison, so as to permit an unbiased estimate of the purported solution's effect was a challenge in the Progresa context as it is elsewhere.

Use of Results. Nobody can guarantee that dependable evidence will be used in policy decisions. In Mexico as in the United States and elsewhere, factors other than dependable evidence—existing law, social and personal values of people who make policy decisions—will and must be brought to bear in decisions. In addition, it is difficult to anticipate, and head off, the work of hypocrites, mischief makers, and saboteurs.

Nonetheless, laying groundwork to enhance the likelihood that dependable evidence from a randomized trial is actually used is important. In the Progresa case, the groundwork was laid in administrative and legislative deliberations on the potential value of research. Discussions of this sort in Mexico are not unrelated to those that led to the initiation of the Tennessee class size experiment in the United States (Ritter and Boruch, 1999) and to the welfare-to-work randomized trials in the United States, Canada, and elsewhere. The process for discussion may involve administrative action, public legislative hearings, judicial rulings, and other activity. An important feature of these processes is that people who hear about the uncertainty in purported solutions would eventually get the evidence from a randomized trial that had a voice in the trial's initiation, the design of the trial, and the tested intervention.

Put aside for a moment the inevitable uncertainty in anticipating the use of any evidence in complex political contexts. Consider the actuality in the Progresa case. The trial's results were positive in the sense that the conditional monetary transfer in the Progresa program worked to increase school enrollment, health clinic attendance, and nutrition. For instance, school enrollment increases in the treatment villages were 20 percent for girls and 10 percent for boys beyond the rates in control villages. Partly on account of these results, the program (under the name Opportunidad) has continued through at least three government administrations in Mexico. This continuation is remarkable. It is in the context of routine terminations of antipoverty programs by each new Mexican government and creation of new ones by each new government. Furthermore, the government expanded the program

to include poor urban areas as well as poor rural villages. Recommendations about extending the program to high schools, made in the report on the trial, were also incorporated into the expanded program.

Human Rights: Ethical Propriety of a Trial. In Mexico's Progresa experiment, families and villages that clearly did not need conditional income transfers did not receive them. This was an ethical and political judgment. Families and villages at the margin and those well below an economic poverty threshold were identified as eligible for assistance based on census and other data and on local surveys. Because the resources for conditional income transfers are scarce, a lottery allocation met a reasonable standard for protecting one kind of people's rights: equitable distribution of resources at a certain level to villages when the total resources are scarce. The lottery-based rollout of interventions across regions of a country and the lottery-based allocation of interventions to different grades in different schools constituted a trial design that satisfies a social standard of ethics, at least for a time.

A Reprise and Observations on Evidence-Based Ethics. These ethical questions and answers are plainly put. Nonetheless, they seem sensible in delimiting the conditions under which we may consider mounting randomized trials to estimate the relative effects of interventions that are purported to reduce poverty, enhance well-being, and advance societies in various ways. They can be used by governmental organizations in the United States. Indeed, these social standards are implicit in congressional and administrative decisions embodied in the No Child Left Behind Act and the creation and activity of the Institute for Education Sciences. They are also implicit in the laws and agencies created to support health research and education research in the United States and the United Kingdom.

Such standards are broadly applicable to deliberations by multinational organizations such as the World Bank, International Monetary Fund, and others. The basic questions can help people in such agencies to decide when a randomized trial might be warranted. They would have to be tailored, of course, to the local standards of organizations, cultures, and countries.

Broad ethical standards are important at the macrosocial level. Their operationalization at the local level, particularly in regard to protecting human rights, is no less important. Institutional Review Board mechanisms, data review boards for ongoing medical trials, advisory boards for new trials, and related arrangements are pertinent. Ethical conflicts, however, may be ostensible or real. Consequently, empirical research that informs decisions about the ethical and legal propriety of randomized trials can be important. The new *Journal of Empirical Research on Human Research Ethics,* for example, is a vehicle for reporting empirical work on ethics. This includes research on how to ensure sensible levels of informed consent in difficult settings (vulnerable groups). The overriding theme is evidence-based ethics decisions.

Advancing the state of the art in this context also includes work on experimental design. For instance, cluster randomized trials involving

facilities, police hot spots, and housing projects as the units of random assignment meet some ethical standards, while random assignment at the individual level may not. Delaying the deployment of an intervention randomly for some people (or entities) and delivering to the remaining randomly assigned units may help satisfy an ethical standard requiring that all units receive the intervention sooner or later. (See Boruch, 1997, for an outline of experimental design approaches to meeting ethical standards.)

Evidence Standards: Fair Comparisons, Unbiased and Biased Estimates of the Effects

Randomized trials, when conducted well, produce statistically unbiased estimates of the relative effects of economic, medical, behavioral, and other social interventions. That is, the random allocation of entities or people to different interventions ensures that there is no systematic difference at the outset among the entities or people assigned to the different interventions. The groups may, of course, differ on account of chance, which conventional statistical methods can take into account.

The guarantee of no initial systematic differences in confounding variables, observable and otherwise, is paramount to the science. Beyond this, the randomization permits making statistically legitimate statements about one's confidence in the estimated magnitude of effects relative to chance variation.

Analyses of data from passive surveys or nonrandomized evaluations or quasi-experiments cannot similarly ensure unbiased estimates of the intervention's relative effect. We cannot ensure unbiased estimates, in the narrow sense of a fair statistical comparison, even when the surveys are conducted well, administrative records are accurate, and analyses of quasi-experimental data are based on thoughtful causal (logic) and econometric models. The risk of misspecified models, including unobserved differences among groups (the omitted variables problem), is often high.

Put aside, for a moment, the intellectual guarantee of fairness of randomized trials. A major evidential issue hinges on the answer to the question, "Do estimates of the effects of interventions based on nonrandomized trials (quasi-experiments) really differ from estimates that are based on randomized trials?" Empirical studies are essential to address this question. The challenge is complicated by, among other factors, considerable variety in quasi-experimental designs.

Glazerman, Levy, and Myers's study (2002) of dozens of trials and quasi-experiments is a case in point. They initiated this work as a test-bed project under the auspices of the Campbell Collaboration (see http://campbellcollaboration.org). Their results should worry those of us who build models based on observational studies and quasi-experimental designs rather than models with fewer assumptions based on randomized trials. In the employment, training, and welfare sectors, they found that the biases

in estimates of the effects of programs based on nonrandomized trials are often substantial and usually cannot be predicted. The variance in estimates based on the nonrandomized trials can also be very large. Although they focus on absolute magnitude of bias, an implication is that one can inadvertently make useless programs look harmful by using some conventional statistical and econometric methods. This indirectly confirms Campbell and Boruch's analytical work (1975) in this vein. Of course, at other times, using different models (designs), one can make useless programs appear to have a positive effect.

The results of Glazerman, Levy, and Myers (2002) are new in some respects but far from new in other respects. Among the earliest examples at hand is Yates and Boyd's agricultural research (described in Snedecor and Cochrane, 1989, pp. 358–359) following World War II. They reported that estimates of the effect of spreading barnyard manure on potato fields, based on regression and related analyses of passive survey data, were large and negative. That is, ordinary least squares regression of crop yield on other variables suggested that the manure resulted in decreased yield. Estimates of the manure's effect based on well-controlled randomized trials, in contrast, were positive and large, with an increased yield. The difference in magnitude and direction of estimates based on the randomized trial versus the nonrandomized trial is possibly related to differential skills of farmers in raising farm animals and spreading manure. Relevant variables were not measured and could not then be included in the regression models.

Examples from criminological research are also at hand (Sherman, 2003). Some make it clear that depending on quasi-experiments can lead one to declare that programs are beneficial when in fact their effects are negative. McCord (2003), for instance, reports on an early illustration, the Cambridge Somerville project, that was thought to be successful partly on account of the intensiveness and context of counseling delinquent youth and the ostensibly positive results of nonrandomized studies. The randomized trial on the program showed that it harmed rather than helped youth. McCord's further illustrations focus on similarly disconcerting disparities between results of randomized and nonrandomized studies of other delinquency programs. These programs, which were thought initially to be successful and were then found to be counterproductive in randomized trials, included the use of court volunteers, group interaction training, specialized activities programs, and Scared Straight programs. In particular, recent and thorough meta-analyses of Scared Straight programs for youth, done under Campbell Collaboration auspices, found that effects of the program were negative or negligible based on the randomized trials and positive based on various nonrandomized studies (Petrosino, Turpin-Petrosino, and Buehler, 2002). The reasons that such differences appear are important to study; no one has done so to date.

In education, the production of methodological studies comparing the results of randomized trials to those of nonrandomized evaluations has not been high. This is partly perhaps because the volume of randomized

controlled trials in education has not been high until recently. Still, there are interesting examples. Wilde and Hollister (2002), for example, compared estimates of the effect of reduced class size based on the Tennessee class size randomized experiment against estimates of effect of reduced class size produced through propensity score matching and related analysis of data from nonrandomized comparison groups. They focused on eleven schools with a sufficient number of kindergartners within school to support analyses of outcome variables such as reading and math achievement test scores. Pretest variables included student-level variables such as gender, race, and school lunch status; classroom teacher-level variables such as teacher's experience; and school-level variables such as system type.

The authors constructed pools of school students for potential comparison from all of the ten schools apart from the target school. For school 1, then, all students in schools 2 through 11 constituted the pool from which comparison students could be drawn. Particular comparison students were then identified and selected using propensity score matching. The matching was intended to ensure a close match between "a child who was treated (reduced class size) in School #1 and a counterpart from schools #2 through #11 who was untreated (regular class size). The estimated effect of reduced class size based on this approach is then the difference in mean test percentile between the kindergarten treatment group (small classes) in the selected school and the comparison group and propensity score matched children (not in small classes) drawn from other schools" (Wilde and Hollister, 2002, p. 8).

Wilde and Hollister (2002) present statistical results in different ways. One bottom-line result is that "for 8 of the 11 schools, the two impact estimates (based on the trial and the propensity score estimates) are statistically different from one another" (p. 13). Their generalization is that "with these data the propensity score matching non-experimental estimates do not do very well in approximating the experimental estimates . . . " (p. 14). The implication is that sophisticated matching methods in quasi-experiments in education may not duplicate results of a randomized trial. Furthermore, understanding when and why disparity occurs is important, inasmuch as evaluators must at times depend on these methods.

Until recently, evaluation of effects of programs in developing countries has depended almost entirely on nonrandomized study designs. Econometric model building, selection models, differences in differences, time series, and so on sufficed. For instance, the World Bank's Web site had only one easily accessible reference to randomized trials before the year 2000. Until recently, the Organization for Economic Cooperation and Development also had done nothing substantial to advance understanding of what works in any active sense of encouraging or mounting randomized trials.

This international scenario has changed partly on account of concerns about credibility and quality of estimates. Governmental emphasis, including the U.S. government's, on evidence-based policy has probably influenced this shift. Some economists at the World Bank, such as Rawlings (2005),

also deserve credit. Still others have produced methodological studies that are used to argue for more attention to randomized trials. The comparison of Glewwe, Kremer, Moulin, and Zitzewitz (2004) of estimates of the effect of using flip charts on Kenyan children's achievement is a remarkable example. There were differences between estimates based on randomized trials versus estimates based on model-based analyses of passive survey data. This begs the question: if one cannot properly model what happens in the absence of flip charts, what indeed can be modeled confidently in the developing country context?

Early examples from medicine have been generated since the 1950s (see Boruch, 1975, for some of these). The Salk vaccine trials of the 1950s, for instance, involved both randomized trials and a parallel set of uniform nonrandomized trials. Estimates of the effect of the vaccine on the incidence of poliomyelitis differed about 30 percent depending on whether one relied on the results from the randomized trials or from the quasi-experiments (Meier, 1972). More recent contributions to this topic in the health care field convey a similarly cautionary message.

For example, Deeks and others (2003) produced one of the most extensive studies of randomized versus nonrandomized trials in the health intervention arena and the way the results differ. They identified and examined meta-analyses that focused on this issue and produced new empirical comparisons based on international multisite trials on stroke and carotid surgery interventions. Furthermore, they avoided the flaws in earlier such comparisons and give succinct summaries of eight deep reviews of empirical studies on the topic. They conclude, "Results of nonrandomized studies sometimes, but not always, differ from results of nonrandomized studies of the same intervention. Nonrandomized studies may still give seriously misleading results when treated and control groups appear similar in key prognostic factors. Standard methods of . . . adjustment do not guarantee removal of bias" (p. iii).

Their review of reviews in their ambit capitalized on searches of ERIC, PsychInfo, and other related social science databases, as well as databases covering health care research.

Lest one think that the physical sciences and engineering are free of biased estimates of effect, recall that one of the reasons for the *Columbia* space shuttle failure has been attributed to the "crater equation" used to estimate a projectile's damage to the shuttle's wing. The equation, or its application, was wrong, to judge from empirical tests carried out as part of the research on *Columbia*'s failure (Chang, 2003).

The physical sciences also offer far simpler arguments. For instance, a member of the audience at a recent World Bank conference offered the advice that economists ought to behave like astronomers. He declared that economists ought to improve on their ability to predict instead of doing randomized trials. Given Thomas Jefferson's taste for calculation and preference for deterministic equations, Jefferson too might have made the same declaration.

No one can disagree with the encouragement to predict better. Nonetheless, evaluators need to recognize that the ability to forecast what would happen in the absence of the intervention is domain specific. For instance, during the early 1970s, the effectiveness of Kevlar in body armor was debatable. To test the effectiveness of the intervention, police researchers draped the cloth over a pig, fired a large-caliber weapon at the pig, and then determined whether there was any bloodshed. The "intervention worked" in that the cloth prevented the bullet from penetrating the pig. That is, no blood was shed. How many control pigs did one need to buttress the relevant causal inference?

One answer then, as now, is "none," provided one is willing to make assumptions. This is because the trajectory of a high-caliber bullet is predictable, assuming that the aim is right and the weapon functions properly. This ability to predict a trajectory, however, is specific to the domain of ballistic equations whose origins date from the seventeenth century if we use Galileo and Newton as benchmarks (Coyne, Heller, and Zycinski, 1985). Governmental and multinational entities cannot wait four hundred years for dependable forecasting equations. Randomized trials provide dependable evidence in the near term.

Less Biased Estimates from Quasi-Experiments and Research Opportunities

One way forward lies in developing an agenda for methodological studies so as to build better empirical understanding of whether, when, and why results of randomized trials differ or do not differ from results of nonrandomized trials. Examples are at hand in which effect sizes in randomized trials and effect sizes in certain related quasi-experiments come close to one another. Including pretests as matching variables or covariates, for instance, seems essential to good bias reduction in the quasi-experiments. (See Glazerman, Levy, and Myers, 2002, for a deep review in the employment and training arenas; Slavin and Lake, 2006, for potentially relevant evidence from studies of elementary mathematics education evaluations; and Shadish, Luellen, and Clark, 2006, on a particularly interesting experiment on randomized versus nonrandomized trials.) Beyond the use of pretests and a small suite of matching variables such as age and gender, however, the role of matching variables within a domain of study (employment, elementary, mathematics) appears not to have been explored well in this context. A federal research agenda might be based on simple but crucial themes of the sort described by Lincoln Moses (1995, 2000), an eminent biostatistician who was also interested in education research: identifying the potentially relevant omitted variables in nonrandomized studies, measuring them correctly, and employing them properly in the statistical model that drives the analysis. Comparisons between results of randomized and nonrandomized studies can provide an important empirical base for scientific understanding.

From the methodological studies of the kind discussed above, the more directive implication is this. Some quasi-experimental approaches and related models will produce estimates of the effect of an intervention that are less biased than other quasi-experimental approaches. Any such approach that includes a pretest as a covariate or matching variable is better than an approach that does not. Before measures (pretests) are bias reducing.

Furthermore, any approach that uses preintervention variables in addition to pretest can also reduce bias. The bias reduction, of course, depends on whether these variables actually influence whether different kinds of people or entities get engaged in particular interventions.

With many potential matching variables, propensity score matching has been shown to be analytically right provided certain assumptions are met (Rosenbaum, 2002; Rubin, 1997). The approach has also been shown to be approximately right in some comparative tests. The GAO report (1994) is a good early example. At times, an ordinary least squares analysis also works. Partly on account of this evidence and analysis, the What Works Clearinghouse in education established a standard that said, roughly, that quasi-experimental evidence is admissible only if there is a pretest and other covariates.

Methodological studies that compare results of randomized and nonrandomized trials are important for at least three reasons:

1. Randomized trials may not be acceptable because they violate human rights.
2. Regardless of social ethics, a randomized trial may not be feasible or desirable at the particular stage of a problem's evolution.
3. Randomized trials may be flawed in design or execution, turning them into nonrandomized studies in which interpretation is more difficult.

Extant methodological studies can be improved in a variety of ways. For instance, Glazerman and his colleagues examined only absolute magnitude of bias in nonrandomized trials. They did not consider scenarios in which the direction of bias may be important. That programs can be made to look harmful when they are merely useless is important to some policy analysts and certainly to some scientists. That some programs can be made to look as if their effects are positive when their actual effects are negligible in some domains is also important to some other policy people and scientists.

It seems sensible to recognize that the direction of statistical biases in nonrandomized trials, including quasi-experiments, can be domain specific. For example, Campbell and Boruch (1975) explained the different ways in which conventional statistical analyses of evidence from evaluations of compensatory education programs could make programs appear harmful when in fact the programs were not effective. A reason for a negative bias in ordinary least squares regression estimates of a program effect, based on

observational data rather than a randomized trial, is that random variables (covariates, the right-hand side of the equation, and the like) are measured poorly. This problem, related in the limit to the omitted variable problem, could easily result in biased estimates of the effects of Head Start, the Comprehensive Employment and Training Act, Youth Employment and Demonstration Projects Act, and other programs when analyses were based on linear models applied to data from nonrandomized trials.

Consider further both uncertainty in estimates and magnitude of bias and how these might be domain specific. For instance, Hedges and Nowell (1995), basing their work on surveys in the United States, and Aubrey Wang (2001), basing hers on cross-country studies in the Third International Mathematics and Science Study (TIMSS), found that boys' performance on achievement tests is more variable than girls'. These studies are based on probability sample surveys. No one has developed a convincing explanation as to why this holds for math across most of thirty countries in the TIMSS. But this domain specificity, gender being the domain, suggests that our ability to forecast for boys is inferior to that for girls, at least at times. The variability in bias, when an estimator is in fact biased, may then also be larger when the target is boys as opposed to girls.

Fraker and Maynard's comparative study of quasi-experimental and randomized trials (1987) is among the few that attended to domain differences. Their work invites an exploration of the prospect that the bias in nonrandomized trials may be lower when the target is mainly women (in the Fraker and Maynard study, those who were receiving Aid to Families with Dependent Children) as opposed to mainly young males (in the same study, youth who were involved in supported-work programs). The complications are obvious. Nonetheless, one implication is that domain may matter a lot in quasi-experimental studies if we are interested in whether estimators are biased and, if so, by how much.

Learning About Randomized Trials and Pertinent Standards of Evidence

Attempting to do empirical comparisons between the results of randomized trials and the results of nonrandomized trials such as quasi-experiments begs a question: How indeed can the thoughtful evaluator identify randomized trials, and locate the results from them, so as to establish a firm standard for judgment? The question is addressed in what follows.

Web-Accessible Registers of Randomized Trials. Until recently, there have been no readily accessible national or international resources for locating randomized trials. The situation changed in 1993 when the international Cochrane Collaboration (see http://cochrane.org) was created to prepare, maintain, and make accessible systematic reviews of studies of effects of health interventions. Randomized trials have been a key ingredient, but not the only one, for these systematic reviews. The Cochrane

electronic library on trials contains more than 350,000 entries. Cochrane has set a remarkable precedent for accumulating and building a knowledge base of this sort.

In 2000, the international Campbell Collaboration was created as the younger sibling to Cochrane to prepare, maintain, and make accessible systematic reviews of studies of the effects of interventions in education, crime and justice, welfare, and other social arenas. Like Cochrane, the Campbell Collaboration depends heavily, but not exclusively, on randomized trials. The Campbell Collaboration's Web-accessible library, the C2 Social, Psychological, Education, and Criminological Trials Register (C2-SPECTR), contains over thirteen thousand entries on randomized and possibly randomized trials. It grows as reports on more trials are located by volunteers who contribute to the collaboration.

For national government agencies and relevant multinational organizations, part of the way forward lies in fostering electronic registers of randomized trials and perhaps also registers of various kinds of nonrandomized trials. Building reliable, continuously improved, and comprehensive registers of this sort is a nontrivial challenge, however. The publication bias problem, for instance, is one that both Campbell and Cochrane have confronted. Both organizations recognize, for instance, that the results of trials are not always made public (especially if the news runs contrary to a particular political view) and that keeping abreast of new trials is a way of ensuring that we can then recognize the suppression of reports as well to keep track of the new studies. For this reason, Campbell is also creating a prospective register of trials as part of C2-SPECTR. That is, grants and contracts for new trials are put into a register.

The more serious problem is the coverage bias in electronic search engines. The search engines on which the World Wide Web depends do not pick out many of the trials. Hand-searching academic journals, for instance, typically yields three times the number of trials identified in a Web-based search (Turner and others, 2003). Surveys of organizations that sponsor or conduct such trials are also essential, partly because many of these do not produce reports in refereed academic journals.

Reporting and Assessment Standards. The design, execution, data analysis, and results of randomized trials are not always reported well in peer-reviewed scientific journals. Until recently, in some sectors, important features of trials have not been reported uniformly, much less completely. In recent years, there have been some improvements.

In the health care arena, for instance, the CONSORT statement provides guidance on what a good report on a randomized trial should contain (Altman and others, 2001). This guidance has been adapted also to reporting on cluster randomized (place randomized, group randomized) trials (Campbell, Elbourne, and Altman, 2004). At least some American Psychological Association journals are using CONSORT to ensure better quality in the reporting of psychological research.

In the education research sector, Mosteller, Nave, and Miech (2004) made a strong case for structured abstracts, a case related to the idea of better and more complete reporting on a trial. Structured abstracts are designed to get beyond the untidiness, idiosyncratic character, nonuniformity, and incompleteness of conventional journal abstracts. In a remarkably brisk response to the initiative taken by Mosteller, Nave, and Miech and by others, the Institute of Education Sciences (IES) of the U.S. Department of Education (ED) tries to generate structured abstracts on all the evaluations (including randomized trials) that the agency sponsors (Institute of Education Sciences, 2006).

Standards for uniform and complete reporting in the text of peer-reviewed journal articles are themselves a vehicle for guiding assessments of the quality of evidence being reported. A report's failure to tell the reader what units were randomized in a trial, for instance, would lead an assessor to regard the report as suspect or undependable.

Evidence grading systems are being developed so as to ensure completeness, uniformity, and transparency in assessment. The IES's What Works Clearinghouse (WWC), for instance, focuses heavily on statistically unbiased estimates produced by randomized trials on education programs, quality in the trial's execution (including attrition issues), and quality and relevance of outcome variables, among other factors. The WWC's detailed coding guides have been revised substantially at least three times since 2000. A similar evidence grading scheme that serves SAMSHA, the National Register of Exemplary Programs and Practices, began revision in 2005.

Organizations such as Cochrane Collaboration and Campbell Collaboration at times lead in development. At times, they follow. Campbell built on Cochrane to produce broad guidelines on reviewing reporting of studies, requiring, for instance, that randomized trials be handled separately from nonrandomized trials (including quasi-experiments). The WWC resources far exceed those of these two voluntary organizations. Consequently, the clearinghouse capitalized on ideas and standards of both and went well beyond Campbell to produce heavily detailed guidance. The Campbell Collaboration will presumably capitalize on these in generating its reviews.

The Society for Prevention Research is among the few professional societies to have produced detailed evidential standards for identifying programs that are efficacious in small scale, effective in larger real-world scale, and ready for dissemination (Flay and others, 2005). These standards are more demanding than others in the genre, partly in the interest of influencing future trials. So, for instance, the guidance requires at least two well-designed and well-executed trials to meet an efficacy standard and two trials on top of these, along with other evidence on generalizability to meet effectiveness standards. Meeting the standard for broad dissemination entails meeting prior standards, good evidence on the interventions costs, and evidence and information on going to scale, such as manualization

and monitoring tools. Constructing these standards led to recognizing that standards for understanding implementation, fidelity of deployment, and replication are weak. Consequently, the society is convening a group to examine these topics.

Scientific standards for assessing evidence or reporting or both, other than the ones mentioned so far, have been developed by various entities. As of 2006, the scientifically oriented ones, apart from the conscientious work of the Campbell Collaboration, the Cochrane Collaboration, the IES WhatWorks Clearinghouse, and some others, included these:

- Child Trends' Synthesis documents on Academic Achievement Programs and Youth Development
- Office of Juvenile Justice and Delinquency Prevention's Blueprints for Violence Prevention
- Centers for Disease Control's Guide to Community Prevention Services
- American Psychological Association's Criteria for Evaluating Treatment Guidelines and the APA *Publication Manual*
- National Institute on Drug Abuse Guidance Preventing Drug Use among Children
- ED's Safe and Drug Free Schools Program
- Coalition for Evidence Based Policy's Effective Programs (2003)

Deeks and others (2003) produced a detailed evaluation of checklists and scales for assessing quality of nonrandomized studies in health-related sectors. Their work pays attention to related themes, including Campbellian threats to internal validity that lead to a biased estimate of an intervention's effect. Deeks and others identified nearly two hundred tools to find that only a third paid sufficient attention to important bias-related factors such as construction and composition of comparison groups, prognostic factor inclusion, and case-mix adjustment.

Rothstein, Sutton, and Borenstein (2005) are among the few who have taken time to address fundamental issues in this arena. The authors in their edited book confront publication bias issues in the scientific literature in the senses of prevention, assessment, and adjustments to the extent possible. The handling is mainly technical in orientation, but it has implications for science policy of course, as this chapter does.

Conclusion

With Thomas Jefferson, we might prefer that error flee more briskly before reasoning and experimentation, but the path is stony and uncertain. Nonetheless, we have made progress using the sturdy vehicles of controlled randomized trials since the middle of the twentieth century.

Over the past decade, invigorated interest in evidence-based policies, programs, and practices has helped to drive up interest in better evidence

on the impact of social interventions, and especially interest in randomized trials. This policy phenomenon can be construed as a delicious opportunity.

Understanding when randomized trials are ethically acceptable and how to design trials so as to be ethically acceptable is important. The Federal Judicial Center Standards are helpful for determining whether randomized trials ought to be considered as a study design. They are insufficient for localized decisions, of course. Federal law in the United States and elsewhere requires the attention of Institutional Review Boards at times. Learning how to design trials so as to meet these local standards will continue to be a challenge in some areas, especially those in which trials have not yet been run. Learning how to generate evidence that routinely informs ethics decisions is also a challenge for a federal government, such as that of the United States, that values human rights in the context of research and the ethical imperative to improve and produce evidence on improvement. This challenge is being met by creative researchers and reported in at least one new journal dedicated to the topic.

Understanding when and why a nonrandomized trial might generate estimates that are similar in magnitude and direction to those generated in a randomized trial is important. If a quasi-experiment suffices, a randomized controlled trial may be unnecessary. Learning empirically about specific domains in which nonrandomized trials suffice, with regard to magnitude or direction, and do not suffice demands special methodological studies that are challenging in design, execution, and analysis. Such work depends on the capacity to predict what happens in the absence of the intervention being tested, including identification of the right variables measured in the right way and incorporated properly in statistical analysis. The direction of biased estimates invites more attention from empirical researchers than it has received.

Understanding how to build cumulative knowledge by building and exploiting retrospective registers of trials and prospective registers of trials is important. Methodological studies of the kind just described would be easier if such registers were developed for this and for other purposes, notably systematic reviews of evidence and meta-analyses. The international Campbell Collaboration in the social sector and the international Cochrane Collaboration in health care are developing such registers. Challenges lie partly in understanding how to make these voluntary efforts sustainable. For federal government agencies and multinational organizations, the challenge lies in deciding whether, when, and how to institutionalize such registers with considerably more resources than voluntary organizations can provide.

Understanding how to enhance the quality of reporting on both randomized trials and nonrandomized trials is important. Absent completeness and uniformity in reporting, we will be seduced by imperfect reports on randomized experiments as we are seduced by quasi-experiments whose features are not made plain. Standards for reporting about randomized trials have been developed, are being evaluated, and are periodically revised.

Standards for reporting on nonrandomized trials, quasi-experiments, are behind these and invite attention. Learning how to apply good standards routinely and how to adjust applications as the standards of evidence are modified will not be easy. Federal government agencies, multinationals, and professional societies would benefit from good research on this topic also.

When we launch programs to redress serious social problems in the United States and in the developing world, we owe it to the people we are theoretically serving to get it right. No less than those of us taking our daily dose of aspirin, they deserve assurance that the interventions they are subject to are effective in improving their life chances.

The world of domestic and international aid is littered with well-intentioned, failed programs, and there is no shortage of projects that are thought to do good but whose value remains uncertain. Randomized trials are essential to establish that these interventions do indeed work.

All of this is in the spirit of fostering thoughtful democratic societies. The goal here is to assist governments and the people they represent in making sound decisions about where to expend social and financial capital.

With Thomas Jefferson, we would prefer error to flee more briskly. Its flight at any pace is an exciting prospect that, for an informed democracy, depends on reason and a society that experiments conscientiously, making honest failures and genuine successes.

References

Altman, D. G., and others. "The Revised CONSORT Statement for Reporting Randomized Trials: Explanation and Elaboration." *Annals of Internal Medicine,* 2001, *134*(8), 663–694.

Boruch, R. F. "Contentions About Randomized Experiments for Planning and Evaluating Social Programs." In R. F. Boruch and H. W. Riecken (eds.), *Experimental Testing of Public Policy*. Boulder, Colo.: Westview Press, 1975.

Boruch, R. F. *Randomized Experiments for Planning and Evaluation: A Practical Guide.* Thousand Oaks, Calif.: Sage, 1997.

Campbell, D. T., and Boruch, R. F. "Making the Case for Randomized Experiments by Considering the Alternatives: Six Ways in Which Quasi-experimental Evaluations in Compensatory Education Tend to Underestimate Effects." In C. A. Bennett and A. A. Lumsdaine (eds.), *Central Issues in Social Program Evaluation.* Orlando, Fla.: Academic Press, 1975.

Campbell, M. K., Elbourne, D. R., and Altman, D. G. "CONSORT Statement: Extension to Cluster Randomized Trials." *British Medical Journal,* 2004, *328,* 702–708.

Chang, K. "Questions Raised on Equation NASA Used on Shuttle Peril." *New York Times,* June 9, 2003, p. 1:38.

Coalition for Evidence Based Policy. *Identifying and Implementing Educational Practices Supported by Rigorous Evidence.* Washington, D.C.: U.S. Department of Education, Institute of Education Sciences, 2003.

Coyne, G. V., Heller, M., and Zycinski, J. "The Galileo Affair: A Meeting of Faith and Science." In *Proceedings of the Cracow Conference,* May 24–27, 1984. Citta Del Vaticana: Specola Vaticana, 1985.

Deeks, J. J., and others. "Evaluating Non Randomized Intervention Studies." *Health Technology Assessment,* 2003, *7*(27), iii–x, 1–173.

Fabian, B. "Jefferson's Notes on Virginia: The Genesis of Query XVII, The Different Religions Received into That State?" *William and Mary Quarterly,* 1955, *12*(1), 124–138.

Federal Judicial Center. *Social Experimentations and the Law*. Washington, D.C.: Federal Judicial Center, 1981.

Flay, B., and others. "Standards of Evidence; Criteria for Efficacy, Effectiveness, and Dissemination." *Prevention Science,* 2005, *6*(3), 151–175.

Fraker, T., and Maynard, R. A. "Evaluating Comparison Group Designs with Employment-Related Programs." *Journal of Human Resources,* 1987, *22,* 194–227.

Glazerman, S., Levy, D., and Myers, D. *Nonexperimental Replications of Social Experiments: A Systematic Review*. Washington, D.C.: Mathematica Policy Research, 2002.

Glewwe, P., Kremer, M., Moulin, S., and Zitzewitz, E. "Retrospective vs. Prospective Analyses of School Inputs: The Case of Flip Charts in Kenya." *Journal of Development Economics,* 2004, *74*(1), 251–268.

Hedges, L., and Nowell, A. "Sex Differences in Mental Test Scores, Variability, Numbers in High Scoring Individuals," *Science,* 1995, *267,* 41–45.

Institute of Education Sciences. *Evaluation Studies of the Evaluation Division, National Center for Education Evaluation and Regional Assistance*. Washington, D.C.: U.S. Department of Education, Mar. 2006.

McCord, J. "Cures That Harm: Unanticipated Outcomes of Crime Prevention Programs." *Annals of the American Academy of Political and Social Science,* 2003, *587,* 16–33.

Meier, P. "The Biggest Public Health Experiment Ever: The 1954 Field Trial of the Salk Poliomyelitis Vaccine." In J. Tanur, F. Mosteller, and W. Kruskal (eds.), *Statistics: A Guide to the Unknown*. San Francisco: Holden Day, 1972.

Moses, L. E. "Measuring Effects Without Randomized Trials? Options, Problems, and Challenges." *Medical Care,* 1995, *33,* 8–14.

Moses, L. E. "A Larger Role for Randomized Trials in Educational Policy Research." Unpublished manuscript, May 18, 2000, with personal communication dated to Boruch. Palo Alto, Calif.: Stanford University and Center for Advanced Studies.

Mosteller, F., Nave, B., and Miech, E. "Why We Need a Structured Abstract in Education Research." *Educational Researcher,* Jan.–Feb. 2004, pp. 29–33.

Parker, S. W., and Teruel, G. M. (2003). "The PROGRESA Trials in Mexico." Paper presented at the Campbell Collaboration Conference on Place-Randomized Trials, Rockefeller Foundation Center, Bellagio, Italy, Nov. 2003.

Parker, S. W., and Teruel, G. M. "Randomization and Social Program Evaluation: The Case of Progresa." *Annals of the American Academy of Political and Social Science,* 2005, *599,* 199–219.

Petrosino, A., Turpin-Petrosino, C., and Buehler, J. "'Scared Straight' and Other Juvenile Awareness Programs For Preventing Juvenile Delinquency." Campbell Collaboration Library, 2002. Retrieved Jan. 24, 2007, from http://campbellcollaboration.org/doc-pdf/ssrpm.pdf.

Rawlings, L. "Operational Reflections on Evaluating Development Programs." In G. K. Pitman, O. N. Feinstein, and G. K. Ingram (eds.), *Evaluating Development Effectiveness: World Bank Series on Evaluation and Development*. New Brunswick, N.J.: Transaction, 2005.

Rawlings, L., and Rubio, G. "Evaluating the Impact of Conditional Cash Transfer Programs: Lessons from Latin America." Unpublished manuscript, Latin American and Caribbean Human Development Department, World Bank, May 2003.

Ritter, G. W., and Boruch, R. F. "The Political and Institutional Origins of a Randomized Controlled Trial on Elementary School Class Size: Tennessee's Project STAR." *Educational Evaluation and Policy Analysis,* 1999, *21*(2), 111–125.

Rosenbaum, P. R. *Observational Studies*. (2nd ed.) New York: Springer Verlag, 2002.

Rothstein, H. R., Sutton, A. J., and Borenstein, M. *Publication Bias in Meta-Analysis: Prevention, Assessment, and Adjustments*. New York: Wiley, 2005.

Rubin, D. "Estimating Causal Effects from Large Data Sets Using Propensity Scores." *Annals of Internal Medicine,* 1997, *127,* 257–294.
Shadish, W. R., Luellen, J. K., and Clark, M. H. "Propensity Scores and Quasi-Experiments." In R. Bootzin (ed.), *Festschrift for Lee Sechrest.* Washington, D.C.: American Psychological Association, 2006.
Sherman, L. W. "Misleading Evidence and Evidence Led Policy." *Annals of the American Academy of Political and Social Sciences,* 2003, *589* (Entire Issue).
Slavin, R. E., and Lake, C. "Effective Programs in Elementary Mathematics: A Best Evidence Synthesis." Baltimore, Md.: Johns Hopkins University, Sept. 2006.
Snedecor, G. W., and Cochrane, W. G. *Statistical Methods.* (8th ed.) Ames: Iowa State University Press, 1989.
Turner, H., and others. "Populating an International Register of Randomized Trials." *Annals of the American Academy of Political and Social Sciences,* 2003, *589,* 203–225.
U.S. Government Accountability Office. *Breast Conservation Versus Mastectomy: Patient Survival in Day to Day Medical Practice and Randomized Studies.* (PEMD-95-9) Washington D.C.: U.S. Government Accountability Office, 1994.
Wang, A. "A Cross National Investigation of Gender Differences in Mean and Variance of Mathematics Achievers of Thirteen-Year-Old Students from a Social Psychological Perspective." Unpublished doctoral dissertation, University of Pennsylvania, 2001.
Wilde, E., and Hollister, R. "How Close Is Close Enough? Testing Nonexperimental Estimates of Impact Against Experimental Estimates of Impact of Education Test Scores as Outcomes." Madison: Institute for Research on Poverty, University of Wisconsin, 2002. Retrieved Jan. 24, 2007, from http://www.irp.wisc.edu/publication/dps/pdfs/dp12402.pdf.

ROBERT BORUCH is University Trustee Chair Professor at the University of Pennsylvania and holds appointments at the Graduate School of Education, the Statistics Department (Wharton School), and the Jerry Lee Center for Criminology.