A TRIALIST’S NOTES ON EVALUATION THEORY AND ROOTS

Robert F. Boruch

Working in the evaluation vineyard is a privilege. It has been profitable in ways that have nothing to do with money. This experience accounts for some of my roots, which are tangled.

THEORY AND EXPERIMENTS: ENGINEERING

Engineering education during the 1960s put little value on “theory,” at least explicitly. As a student at Stevens Institute of Technology, I (and others) aimed to build a structure that could stand up in a hostile environment. A bridge that would carry people safely. A deck cannon that would not explode and kill its users. A sheet of plastic that would be uniform in quality and perform well.

The theory was implicit in the equations that engineering students and engineers used to design the equipment. Those equations took time to understand. But some of the equations (theory) were imperfect.

I had depended on equations to design the extrusion barrels (cannon-like devices) to produce plastic sheeting, and learned that the barrels at times failed. I learned further that the equations were developed during World War II for the
design of naval deck cannons and that those cannons failed at times also. This was a secret of the war. So were the experiments on the cannons and the failure of the equations. For a contemporary example of the same kind of learning, see Kenneth Chang’s article (2003) on the Crater equations used to estimate the effect of low-mass, high-velocity debris hitting the Columbia shuttle. The more general understanding is that the mathematical models can be wrong, as well as right, in engineering, econometrics, or education. Evidence from experiments can trump these models.

STATISTICS, MODELING, AND THEORY: GRADUATE EDUCATION

As a graduate student at Iowa State University, statistical theory came as a revelation. Coherent, beautiful. A Poisson distribution, Student’s, and so on could accurately characterize some realities. More pertinent here, the idea of a statistical model as rudimentary theory of how one variable might be correlated with, or even caused by, another variable and could also account for ignorance (a packet of random error) was simplistic, but beautiful.

Fit of the model to the data was, and is, a way to understand whether the theory matches a reality. I worked my way through early versions of latent trait theory, structural modeling, and path analysis, publishing some decent papers in the process. Statistical models of passive observational data helped build plausible story lines, as indeed they are supposed to do. Leroy Wolins was splendid in teaching this. Indirectly, others have done so also, and some get Nobel prizes for it (e.g., Jim Heckman).

Simpler theory and a basic interest in learning about cause and effect in a less equivocal way than the modeling of observational data permits drove me to randomized trials. Results of these trials are represented by rudimentary statistical models that permit unbiased estimates of parameters that reflect cause—a relative difference between the effects of two or more interventions when the interventions are assigned randomly to individuals or entities.

This was a sturdy and reliable tool for handling a part of all science—discovery of a cause-effect relation—in a way that yielded less equivocal results than its competitors. My early mentors at Iowa State University, notably Leroy Wolins and Oscar Kempthorne, a protégé of Sir Ronald Fisher, deserve credit for teaching us about this.
POSTGRADUATE EDUCATION AND EXPERIENCE

The Social Science Research Council (SSRC), under Henry Riecken’s brief leadership, contributed remarkably to advancing understanding. Its Committee on Social Experimentation, for instance, included economists, sociologists, psychologists, statisticians, and education researchers (Riecken & Boruch, 1974). That committee and its advisors taught me more about the why and how of randomized trials.

SSRC itself helped many scholars to understand that there are worlds beyond statistical models and measurement, including history. Connecting the worlds was an even greater joy (Boruch, 1984).

Part of the virtue in this SSRC Committee effort lay in bringing ideas about randomized trials out of the laboratories of psychology and the agricultural fields and into the social sector. It assembled fine people with different tastes: Don Campbell in psychology and Al Rees in economics, among them, to focus on the topic. As an SSRC research associate, I learned a lot from Peter Rossi, Harold Watts, Carol Weiss, and Rob Hollister, who consulted for the committee and learned how to get beyond some of their ideas at times.

I got to Northwestern as a research associate on one of Don Campbell’s NSF grants. Don’s behavior taught me to understand the virtues of the young and industrious, and to hire them, work with them, and admire their efforts often. I walked through Northwestern’s front door when an assistant professorship developed. It was mighty easy to learn from Campbell about how to think about science, questions, and methods appropriate for addressing the latter. Fine senior colleagues in disparate branches of psychology—Ken Howard, Benton Underwood, and Lee Sechrest—were creative, smart, and alarmingly productive, and generous in sharing ideas.

Closer to my age, people such as Dick Bootzin, Tom Cook, Dick Berk, Chuck Manski, Bill Revelle, and Paul Wortman educated me also. Northwestern’s postdoctoral and predoctoral programs are a fine memory. If postdoctoral fellows are part of an invisible college of methodologists who continue to contribute to understanding, Laura Leviton, Dave Cordray, Will Shadish, Georgine Pion, Bill Trochim, Terry Hedrick, and Bob St. Pierre are among them. Students with whom I’ve published and students on whose dissertation committee I served have also been my tutors. Joe Cecil, Hernando Gomez, Bill Trochim, Mel Mark, John Soderstrom, Chip Reichardt, and Mike Dennis are among them. It was a delight, as professor, to work with Bruce
Spencer, Myer Dwass, Sandy Zabell, and others to create a Statistics Department at Northwestern, but not a delight to fail to get the president’s support for a Center on Social Experimentation.

The U.S. General Accounting Office (GAO), from the 1970s through the 1990s, became another invisible college. Comptroller General Elmer Staats created a Program Evaluation and Methodology Division (PEMD), which continued until near the end of Comptroller General Chuck Bowsher’s term. PEMD elevated the GAO and congressional understanding of evidence in all social sectors and produced reports based on that evidence that still constitute great case studies. PEMD helped to attract a fine stable of talent to the GAO. Other members of the PEMD Advisory Group included Dick Light, Harry Hatry, and people who were equally alarming in their industry, brains, and capacity to do some good. It was also a privilege to serve the GAO during 1976 to 1996 as a member of the Comptroller General’s Research and Education Advisory Panel. Other colleagues learned along with me.

At times, committees of the National Academy of Sciences can be fine opportunities to learn to develop different roots over time. They included the Committee on Concorde noise level, where we learned how FAA (at the time) could do surveys that made a fashionable airplane’s noise level appear deceptively low.

Mosteller, a member of the Committee, and Angus Campbell, its chair, were a delight in teaching some of us how to ask questions and how to reserve judgment until the evidence is in. That airplane was noisy. The Committee on Youth Employment Programs, chaired by Rob Hollister, helped to reshape research on employment and training programs in the United States. They included the committee on AIDS Prevention, where Harry Hatry, Lincoln Moses, Robyn Dawes, Charles Turner, and I learned a lot about the different ways to estimate effects of such programs and the political, managerial, and statistical problems that each engenders. They include the Committee on Scientific Principles in Education Research, chaired by Richard Shavelson. Two of its members, Eric Hanushek and Jack Fletcher, were particularly important in helping me to understand how to justify and argue for randomized trials. The chair and committee educated me about the context for trials and larger issues of science.

The Spouse Assault Replication Program, sponsored by the National Institute of Justice, aimed to understand whether arresting misdemeanor offenders would affect recidivism. It was easy to agree to be part of a program
review team or randomized trials in this arena. But I had to buy two bulletproof vests, drive with cops into the wee hours, and take the risk at age 45 or so of getting my bones broken or my body parts shot. The substantive theory underlying these trials was interesting: specific versus general deterrence. The evidential theory was interesting. In all this, Al Riess, elder statesman in criminology, Larry Sherman, Kinley Larntz, Al Andrews, and Joel Garner, seasoned bureaucrat, were the best of colleagues.

At the U.S. Department of Education, Al Ginsburg developed an Evaluation Review Panel (ERP) that had no official standing. Its members were nominated by professional organizations, but it could not be counted as a prestigious government committee, which operates under certain federal regulations. The ERP was a good idea for its time and included some able people. Ginsburg and members of ERP, the staff at what was then called the Planning and Evaluation Service, were excellent in trying to learn from one another.

Since 1999, the international Campbell Collaboration has occupied my attention. It comprises many able colleagues and mentors. Its main aim is to produce systematic reviews of studies of the effects of interventions in crime and justice, welfare, and education (http://campbellcollaboration.org). Its older sibling is the Cochrane Collaboration in health care (http://cochrane.org).

The emphasis in both is on randomized trials that produce fair comparisons. Both the Campbell Collaboration and the Cochrane Collaboration engage people who are interested in such comparisons. They are eager to not fool themselves, and to assure that others are not fooled. My mentors and colleagues in this initiative have included Sir Iain Chalmers, Frederick Mosteller, Harris Cooper, Larry Hedges, Will Shadish, David Myers, Peter Rossi, and many others, especially Dorothy de Moya (Executive Officer) and steering group members such as Dennis Cheek, Phil Davies, and Haluk Soydan.

The Campbell Collaboration people, from 1999 to 2002, were excellent and generous in trying to achieve the Collaboration’s aims. What is more interesting is that governments are also trying to achieve the same ends. Government agencies do this with substantial and stable resources, which the Campbell Collaboration has not had. The U.S. Education Department’s What Works Clearinghouse (WWC), for instance, aims to do a better job and is being assisted by Campbell (http://www.w-w-c.org). The WWC effort is important in the United States for education, and it will be both painful and a pleasure to learn from colleagues who lead the effort in government, notably...
Russ Whitehurst; colleagues from an earlier time, such as Phoebe Cottingham; and the stunningly able scholars and managers contracting organizations. Becki Herman, Stephanie Cronen, Bill Morrill, Steve Fleischman, Dorothy de Moya, and Jeff Valentine are among the tutors.

HOW TO ESTIMATE THE EFFECTS OF INTERVENTIONS

Given these roots, how might I go about doing an evaluation? The steps here are fundamental to any science. I focus on randomized trials for the simple reason that my roots, though tangled, have led me to make contributions often in this arena. The details are given in a book I wrote (Boruch, 1997), which is called excellent by at least one productive colleague whom I admire (Orr, 1999), and in a book containing excellent papers edited by Fred Mosteller and me (Mosteller & Boruch, 2002), whose authors are excellent tutors, with whom neither I nor others can always agree.

First, identify the right questions among the many questions that the evaluator might address. The right question depends on stage of the evaluation and stage of development of the program. The right question drives the methods used to generate the evidence that helps to answer the question.

Suppose, for example, that the problem of first-graders’ understanding of math is not well understood. Doing surveys, or process-oriented research, and diagnostic tests will help to illuminate the problem. Controlled trials are not designed to provide such information. They presume the existence of interventions that are supposed to resolve the problem.

If the problem lies in deploying programs, methods and research from management, public administration, and government, from anthropology, measurement, and operations research can be brought to bear. Experts in randomized trials often cannot address the question of how to implement or deploy the program. However, they can help address questions about the relative effects of deployment efforts.

Third, if the problem is reasonably well understood and a coherent intervention can be deployed to address the problem, a natural question to pose is whether the intervention works relative to some alternative. The alternative may be a control condition, or it may be another intervention that is supposed to work well. The main point is that a fair comparison is essential if the question is, “What works better?”
If this indeed is the right question, determining whether and how to run a randomized trial is important. The decision should rest on basic conditions for the ethical propriety of trials.

If the problem is severe, then trials are better justified. Interfering with people’s lives to solve a trivial problem is unwarranted.

If the solutions to the problem are debatable, then trials may be warranted. Testing the Salk vaccine again would be unethical, given the evidence available from earlier trials. Determining whether nonrandomized trials yield more equivocal evidence than randomized trials in the particular instance has to be based on theory and data. Designing a test of the claims that body armor works, in the sense of preventing a bullet’s penetration, can rely on substantial data and theory about what happens in the absence of armor. Recall the earlier remarks about deterministic engineering models and statistical modeling of observational data.

On the other hand, one cannot rely on considerable evidence and theory in the design of a test of a new crime prevention program, such as intensive patrol of police hot spots, or of certain new curriculum programs for math. We cannot forecast accurately what would happen to people in the absence of the program. Consequently, developing a fair comparison group through randomization is justified. Nonrandomized trials yield evidence that contains fewer guarantees of fairness in comparison, that is, unbiased estimates of relative effect, than do randomized trials. This is a major scientific justification for considering such trials.

Of course, the evaluator ought to expect results to be used, otherwise the trial is not warranted. Anticipating use is hard. The risk of nonuse is reduced considerably, and the likelihood of use enhanced, if potential users and stakeholders are brought into the trial’s design. All of the major trials done in the United States nowadays, and many of the small- or medium-scale trials, routinely depend on advisory groups at the front end and throughout the trial so as to assure useful results and to assure that the design and execution of the trial are good. I likewise rely on advisory groups, in this and in other arenas.

Knowing how to develop the trust and partnerships needed to mount randomized trials is not a virtue possessed by many. In the best of cases, one can rely on people who know how to do this and do it well. They are partners—teachers, police officers, and so on—whose cooperation is necessary in the trials. Investing in these people, when they can be trusted and can trust the
evaluator, is part of better understanding that leads to better trials. Every good trialist I’ve mentioned here develops the trust. So should other aspiring trialists.

REFERENCES


