Social Experiments: Education[1]

Robert F. Boruch
Northwestern University

1. Introduction

On June 30, 1980 a group at Northwestern University submitted a report to the Department of Education on evaluation of federally supported education programs. The report was mandated under the 1978 Educational Amendments and covers evaluations at the federal, state, and local levels. Congresswoman Elizabeth Holtzman of New York introduced the bill that required the study. Her interest lay in understanding the product of the government's $30-40 million annual investment in evaluation. The enterprise was designed then to answer questions about why evaluations are undertaken, who does them, how well they are done, and how the results of evaluation are used.[2]

We found lots of interesting problems, not the least being lexical promiscuity: A federal director of research, for instance, announced that he did no evaluations, and fifteen minutes later his boss told us that everything in his shop was evaluation.[3] We encountered a Congressional staffer who announced that evaluations are not used and are silly, only to find later in the conversations that evaluation reports were used to guide Congressional hearings and that the standard for "silliness" is akin to the one dowsers use to find water.

My remarks here focus on another aspect of the report: our recommendations on the use of randomized experiments in evaluating education programs. This includes settings in which children, classrooms or entire schools are randomly assigned to one of two or more methods of improving education to estimate their relative effects. The design is a sturdy device for evaluating outcomes. But as many of you know, it is rather difficult to use in social settings.

2. The Recommendation to Congress

The report made six major recommendations to the Department and to the Congress. The recommendation bearing directly on experiments was:

"Good evaluation designs...are not used often, partly because innovations are planned independent of evaluation. We recommend that pilot testing be undertaken before new programs or variations are adopted and that the introduction of new programs be staged so that good designs can be exploited. Further we recommend that higher quality evaluation designs, especially randomized experiments be authorized explicitly in law for testing new programs, new variations on existing programs, and new program components."

The rationale for suggesting pilot tests in education is no different from the rationale in medicine, energy production, and other areas. Higher quality evaluations are much more feasible before the program is adopted at the national level. Better evaluation designs can be employed, conclusions are less likely to be ambiguous, and political-institutional constraints are less likely to be severe. The introduction of new programs can be staged so that earlier stages constitute pilot tests for the later ones.

This is terribly simple, even mundane. But recognize that in current political discussion of the proposed Youth Initiatives Program, for instance, an enterprise whose costs may exceed $850 million per year, there has been no formal attention to pilot testing or staged introduction of the program. The Title I compensatory education program evolved in the same way ten years ago. What we know about its effects is still meager on that account. Reiterating the notion that massive new programs ought to be pilot tested seems to be warranted simply because it is not yet a common practice.

The second part of the recommendation, concerning higher quality evaluation designs, is based on the presumption that we won't learn how to bring about detectable change in the performance of children or schools without more conscientiously designed tests. Its justification lies partly in the miserable quality of designs used in the field. It is discouragingly easy to find, for example, testimony offered to Congressional Committees, in which a Title I program is declared to be a success by a state legislator because "test scores went up." Very little attention is dedicated at the local level to competing explanations for gains in achievement, e.g. normal growth apart from special programs. A kind of benign hypocrisy characterizes the business: Gains are attributed to the program publicly but privately there's some admission of doubt. The same is not true for recent major federal evaluations, however. A few have been nothing if not blunt in reporting and zealous in their search for plausible competing explanations.

We believe explicit statutory provisions permitting randomized designs are essential for two reasons. First, such designs are best in principle if our standard is scientific evidence and that should be recognized publicly. Second, we expect the authorization to facilitate the local evaluator's efforts to conduct decent tests in the face of resistance or indifference of administrators, teachers, parents, and so on.

3. Feasibility and Appropriateness

The usefulness of randomized tests in principle is generally not at issue in professional discussions about evaluating new education proprograms. There is agreement that when experiments are conducted properly, orthodox theory guarantees that long run estimates of effects will be unbiased. The conditions under which

they can be or should be employed is more debatable.

Some squabbles concern the idea that randomized experiments are rarely feasible in field settings. Rareness and feasibility are, however, infrequently specified by government policy groups or by individual analysts. It is true that although the design is not new, its application in evaluating educational and other social programs is relatively novel. The novelty does not establish lack of feasibility and a notable if not large number of field tests have been mounted.

The most recent examples include evaluations of parts of the Emergency School Aid Act, by the Systems Development Corporation, some career education programs supported by the National Institute of Education, Middle Start programs run at Oberlin College for high school students with intellectual promise, educational T.V. programs in preschool education such as Sesame Street, radio-based mathematics instruction programs run in Nicaragua, and even grade retention. Wayne State University staffers have executed remarkable tests to establish the effectiveness of preventive health care for preschoolers with emotional problems and enrolled in Detroit's Title I programs. The Cali, Colombia tests on programs for malnourished and educationally deprived children were a milestone experiment for the developing countries. See Boruch, McSweeney, and Soderstrom (1978) for a bibliography.

Judging from precedent then, bald claims that it's impossible to assign individuals or other units randomly to programs for the sake of fair estimates of program effects are unwarranted. Precedent is persuasive only in the crudest sense, of course. It implies that what has been done, might be done again. But it may be immaterial to the situation at hand.

Pilot Tests of Randomized Experiments

We believe that pilot tests of large-scale field experiments can yield more direct evidence on their feasibility. That is, small experiments prior to the main field experiments can help to identify anticipated problems in the field and to resolve them. Early experiments can inform the conduct of later ones where a sequential plan can be exploited.[4]

The main justification for considering such pikot tests is to get more direct evidence on feasibility than history can offer, to identify problems which cannot be anticipated, to resolve anticipatable problems before the main effort. That there are lots of problems in mounting randomized tests is clear. They fail to be successfully implemented in education as in medicine, law-enforcement, and other areas because the randomization is corrupted, because the programs are not implemented as advertised, because of attrition and for other reasons. Further, a good deal of post facto criticism is directed at execution of designs rather than the designs themselves. To the extent that pilot tests can help avoid unnecessary argument after the experiment's completion, they are sensible.

Appropriateness and Feasibility: Conditions

Precedent won't always be available to guide decisions about experiments. And there will often be little time for pilot tests or exploratory work on feasibility. So it is reasonable to educe general conditions under which experiments might be mounted. We propose five conditions that ought to be considered in the decision to do randomized tests based on earlier work: (a) information needs, (b) randomization equity and mechanics, (c) variations, (d) statistical issues, and (e) law.

(a) Information Needs. The questions that ought to be asked about a proposed experiment include: What will be learned? How will the information be used? What are the alternatives? Considering the first, there's not much point in the investment of scarce evaluation resources in an experiment if experts expect the effects to be trivial or the size of the effect to be very large, if accurate estimates are not critical, or if the effects are already well established. These matters do appear to have received attention in federal committees on educational evaluation. We merely reiterate them here. The second question is more demanding in that major evaluations are supposed to be used in policy decisions. If there's not much chance of this, then the experiment is pointless for many policy-makers. The factors which vitiate interest here include a sturdy indifference to evidence that characterizes many societies, the embarrassment or related difficulties engendered by unflattering evidence (e.g Don't ask questions if you don't really want the answers), and the probable management difficulty of using the information. The persistent failure to lay out decisions which might be taken based on possible outcomes exacerbates the problem of answering the question. Our inability to track utilization well and to predict its occurrence makes matters even more difficult.

The matter of alternatives is a reminder that other kinds of evaluation exist, and the decision we make ought to recognize them. Goals of some programs are overblown, operations often ambiguous. For some analysts, this is sufficient to give outcome evaluations, such as randomized field tests, low priority. More generally, establishing who needs the service, whether they are served, how well they're served in an operational sense, and how much services cost, may all be more important than estimating effects. The point is that options other than outcome evaluations exist. Identifying them may not be easy but the choice ought to be explicit.

(b) Randomization. The second condition bears on feasibility of experiments and the notion of equity. Where there is an oversupply of eligible

recipients for a scarce resource--program services--then randomized assignment of children or other units to the resource seems fair. So, for instance, Vancouver's crisis intervention program for youthful status offenders affords equal opportunity to eligible recipients. Since all could not be accommodated well with available program resources and they are all equally eligible, they are randomly assigned to program or control conditions. Experts such as Cook and Campbell argue that randomized experiments are most likely to be carried out successfully when the boon, real or imagined, is in short supply, and the demand for the boon is high. This rationale dovetails neatly with normal managerial constraints. That is, new programs cannot be emplaced all at once and all eligible candidates cannot be served at once despite the aspirations of program advocates. Experiments can then be designed to capitalize on staged introduction of services.

The mechanics of randomization are no less important than the equity based rationale of course. Field studies in education, medicine, training, meteorology, and other areas have failed because the randomized assignment was subverted. The key to the matter seems to be complete control of the randomization process and prior agreements to adhere to the result. Neither is easy to obtain in educational settings though there's no evidence that it's more difficult in medicine or other areas.

(c) Variations. The third condition concerns settings in which it is politically impossible to assign individuals or other units randomly to control conditions despite the fact that we know nothing about whether a program works. The ethical, moral, and economic justification for experimenting may be quite irrelevant. In such instances, it is often possible to ameliorate difficulties by comparing program variations against one another, rather than comparing a novel program to an existing one or to no program at all. A "No program" control condition may be an unacceptable political option if the program fails anyway. The most we can reasonably expect then is to choose the invented variation or component which works best for the investment.

The idea of testing variations or components rather than testing a program against a control condition is a compromise, perhaps a cowardly one. But I believe that getting some decent information on a subordinate question, such as which variation or component works best, is better than getting no information at all on, the main one--what are the effects of the program. And the idea is generalizable. In particular, for ongoing programs of the motherhood and apple pie genre, it seems sensible to think in terms of randomized assignment to new variations or components to discover more effective or cheaper versions of the program. I do not believe this strategy has been routinely employed by any major ongoing federal education program. It is not common in any agencies except the Census Bureau, and NCHS, where randomized tests are periodically run to understand better methods of doing surveys.

(d) Statistical Matters. There are some topics apart from randomization which need to be addressed by statisticians in the decision to experiment. The topics are not much different in this arena than in medical trials. They include statistical power, which is often computed but has rarely been reported in recent large scale evaluations, and has rarely been computed in the small scale ones. They include probable attrition of individuals assigned to various treatment conditions and how to cope with it and its effects on estimation. They include decisions about unit of randomization and unit of analysis and how that decision will affect subsequent inferences. At least crude aspects of these matters ought to be addressed in the decision to evaluate.

Several issues are distinct from those normally considered by statisticians in medical or agricultural research. Social programs are normally complex and assaying their delivery in numerical terms is difficult. Nonetheless it's essential that we understand how to measure implementation if the experiment is to be informative. Similarly, information about the sensitivity of the response variables is often ambiguous. That makes power computations difficult, control over quality of measurement essential, and estimates of reliability exceedingly helpful. These receive little concerted attention even in large scale studies.

(e) Law. The final criterion which seems to be important concerns the legality of randomized field tests. Not much has been written by legal scholars on the topic partly because they don't know much about it. But that is changing too. So, for instance, the Federal Judicial Center, the research arm of the Supreme Court, now has a Committee on Social Experimentation which is issuing a policy statement on what posture judges should take in looking at experiments which assay effectiveness of judicial changes. There have also been a few pertinent court decisions. For example, in Aguayo v. Richardson and California Welfare Rights Organization v Richardson, the use of randomized experiments in assessing the welfare programs were challenged and the challenges were dismissed by the court. Laws which specify the legitimacy of randomized experiments are rather scarce however, and that is one reason why we recommended more explicit statutes.

4. Other Pertinent Recommendations

Several other recommendations in the report are indirectly pertinent to randomized tests.

We urged the Congress, for instance, to be more direct in its demands for information where directness is possible. In particular, we suggested that laws request information about who is served, need for service, nature of service, and/

or effect of service, rather than just asking for an "evaluation." And we've recommended more regular discussion between agency staff with evaluation responsibility and Congressional staff with related interests to clarify the questions and identify the unanswerable ones. One part of the rationale for this is to assure that outcome evaluations are mandated clearly, and that their difficulty is understood.

A second recommendation called for routine balanced critique of major federal program evaluations and of a sample of locally conducted evaluations. Part of the rationale for this is to identify good and poor quality evaluations as such, and to encourage better quality tests.

The worst of the poor reports are ghastly. For instance, we found reports announcing dozens of F ratios of 9000 or more, references to t tests as being "convenient because their mean is 10 and variance 20," to ".001 as the highest probability one could achieve" in a statistical test, and so on. The recommendation to apply better methods won't help here. Critiques and technical assistance might.

A third recommendation asked that evaluation capabilities at the local and state level be assayed before new evaluation demands are imposed. It was based on the finding that capabilities and demands vary enormously. To the extent that evaluations involve estimating program effects, then the expertise required by the demand for such evaluation ought to be recognized. Our recommendation to provide technical assistance is predicated on such capability assessments.

Though it's focused on local and state agencies, this recommendation applies to some federal operating agencies as well. We are aware of few high quality methodological projects in bilingual education grant programs, for instance. Moreover, the transformation to a Department of Education, and the lodging of the evaluation unit in the Office of the Assistant Secretary for Management, has been accompanied by a dramatic loss of able staff from the unit. Our own report did not handle this problem partly because we were not equipped to examine it, partly because the organization changed late in the course of the study. The loss of well-trained staff does make it difficult to see how any of our recommendations can be activated.

One of our final recommendations urges the Congress not to adopt uniform standards for evaluation in law. But it does encourage adherence to sensible guidelines in major evaluations. Guidelines have been produced, for example, by the U.S. General Accounting Office and by independent professional organizations. And based on our findings of poor quality in some field settings, they deserve attention. The GAO guidelines incidentally are similar in more than a few respects to the advice Mosteller, Gilbert, and McPeek (1980) offer to editors in review of medical journal articles.

5. Some Topics that Require More Attention from Statisticians

Judging from our own work and from others' research, several aspects of educational experiments deserve more attention from statisticians. These topics are not confined to education of course. See Mosteller, Gilbert and McPeek (1980) for a different selection.

Cost Benefit Analysis of Experiments

The absence of formal cost-benefit analyses of outcome evaluations, including experiments, is glaring. Part of the problem lies in defining the benefit. Benefits are often not clear unless the evaluation is used, and "use" is often neither well documented nor well understood. An experiment's finding that a new program was unsuccessful, for instance, might imply an increased budget by way of salvaging a product in public demand, or a decreased budget in the interest of spending resources on more promising projects. Narrowly defined "use" may then be misleading and there is some risk of defining use so broadly as to make it meaningless. The difficulty of specifying political decisions beforehand exacerbates the problem. There are a few good illustrations of orderly accounting of this sort, in day care, fertility control, and meterology. These suggest that there is indeed a class of evaluations that is amenable to cost-benefit analysis but more intellectual attention is warranted.

Parochialism and Statistical Methods

The statistician's normal focus is on statistical methods and that is fine. But it seems to me that statisticians have a responsibility to broaden their view by recognizing that in order to use good statistical methods, one also may have to develop better methods of other kinds.

For example, in Social Experimentation (Riecken et al, 1974), we stressed several classes of problems only one of which was statistical or scientific. The other classes include managerial difficulties, political-institutional problems, ethical, and legal dilemmas. The idea here is that any of these classes of problems can affect the quality of a social experiment. And one needs to develop managerial solutions to management problems, legal or procedural solutions to legal problems, and alternative methods of resolving political-institutional difficulties in order to exploit good statistical methods of estimating the effects of social programs.

The point is that in order to use good statistical methods, one has to have available nonstatistical methods to solve problems. It may be presumptuous to argue that statisticians must learn about nonstatistical methods in order to better learn how to exploit their designs. But it's hard to see how good designs can be exploited better without that education.

Experiments in Social Settings

Statistical methods are transferable in principle across substantive areas. More important, the transfer always engenders new and interesting problems.

So for instance, employing randomized experiments in social settings involves a variety of problems which are not treated in textbooks such as Cox's, Kempthorne's, or other classics. Those problems include the fact that treatments are rarely delivered as advertised, that they are not "fixed" in the same sense that treatments are fixed in the chemical sciences or biology, the fact that the response variables are often measured on scales which do not have (roughly speaking) equal intervals and so estimates of program effect may be biased by floor and ceiling effects, and so on. There is, for example, an awesome array of statistical problems, as well as managerial ones, implicit in understanding criminal justice "treatments" from the structural level down to the individual level (Sechrest and Redner, 1979), and similar problems emerge in education. The point is that these are important problems which have not been well articulated in the orthodox literature in randomized experiments. Bringing randomized design into the social program evaluation area means we have to solve them, and this may lead to innovation.

Incidentally, the occurrence of special problems in transferring methodology from one area to another seems to me to be a matter of degree rather than an all-or-nothing occurrence. Some problems, such as failure to implement treatment perfectly, are more severe in the social sector than they are elsewhere. But they have occurred in agriculture and the hard sciences. One of the purposes of our working papers on comparative affects of social program evaluation was to lay out some of the common problems.

Design for Early, Interim, and Late Results

Many evaluators believe that experiments depend heavily on planners' willingness to wait for results before initiating program changes, since stability is required. I agree with the premise. But it does suggest several other options that ought to be examined.

The impatience implied here, for example, is a two-way street. It is in some measure justified and does appear to be persistent. Consequently, it's reasonable to argue that, as research designers, we should develop plans which always provide supplementary information early in a long-term study, perhaps better administrative information during the study, and that we be able to identify changes which occur in midstream which theoretically have no notable effect on outcome in the experiment. More generally, it behooves us to invent coherent theory to cover the need for information during an experiment rather than only after the experiment's termination. We have no such theory now, though the state of the art is developing.

The other side of this street concerns unwarranted impatience. Most social problems are chronic, and they are resolved in small steps over a long time period despite innovative social programs dedicated to their elimination, rhetoric notwithstanding. For these and other reasons, Congress ought to be reminded discreetly about ingenuous expectations. Some Congressmen are educable, in this respect, I'm told, and I know that some legislative assistants are well informed and thoughtful about the matter. How we get the education done for the thoughtless clients is not clear, but the effort ought to be made. Turnover in client groups makes the problem chronic.

Coupling Randomized Experiments and Nonrandomized Tests

The technology of designing randomized tests has not developed independently from design of nonrandomized tests such as quasi-experiments or time series analysis. But the separation is sufficient to prevent statisticians from thinking about both in design of evaluations. There are some good reasons for thinking in terms of both, for coupling approaches when the opportunity arises.

Part of our suspicion of nonrandomized tests, for example, is based on our ignorance about mispecification and competing explanations for what caused the effect. Yet, it is possible in principle to design simultaneous randomized and nonrandomized tests to estimate biases based on the latter in evaluating a program at hand, and to accumulate empirical estimates of bias of nonrandomized tests for future evaluations where randomization is not possible. Illustrations of the approach are forced by circumstance rather than designed this way at the start, e.g. the Salk polio vaccine trials.

More generally, it is not clear how to design for sequential trials in which randomized tests alternate with nonrandom tests in the interest of accommodating guinea pig effects and other problems engendered by experiments but may not be engendered in nonrandomized tests. It is not clear how one ought to design experiments in anticipation of internal analyses which ignore the randomization category at least partly, but which are also used to inform policy. Nor is it clear how to decide whether combining estimates of effect is warranted, and how to combine, when confronted with a set of independent randomized and nonrandomized field tests, a problem which Pillemer and Light (1979) are assaulting for the case of randomized experiments.

Federal Evaluation Policy and Federal Statistical Policy

No effort has been made yet by any private or public agency to link federal statistical policy (Federal Statistical Project Staff, 1980) with developing federal evaluation policy (Boruch,

et al, 1980). Their separate development, inevitable perhaps in any new and complicated arena of human enterprise, obscures some important points of contact. Both direct attention to statistical design issues, though statistical policy emphasizes sampling matters and evaluation policy is concerned more with planning experiments and quasi-experiments. Both concern themselves with validity of information though the stress on the topic appears to be greater in evaluation partly because measurement error induces alarming biases in estimates of program effects when designs are not randomized. Both involve privacy/confidentiality problems, though evaluation must accommodate other problems as well, e.g. ethical problems of random assignment. The distinctive feature of some evaluative research, qualitative case study, is not recognized in federal statistical policy and that's something of a shame given the ecumenical statistician's interest in exploiting such information in building better surveys. The two cultures provincialism is less evident in evaluation policy.

Both fronts are subject to similar problems of course. Building a good federal statute on experimentation is as difficult as building one on surveys, and we lack consolidation of the little experience we have. Clearances problems--the Federal Educational Data Acquisition Council, the OMB process, the non-government Council for Educational Information Systems--are severe, complicated by bureaucratic warfare and professional incompetence. And these too are a proper target for remedial policy.

Footnotes

[1]Background research for this paper on educational evaluation has been supported by the National Institute of Education (Grant NIE-G-79-0128). The National Science Foundation provided support for related work in the social sciences (NSF-DAR-7820374). Substantive work on the Holtzman project was supported in 1979 by OED-300-79-0467. Opinions registered in the paper are not necessarily consistent with agency policy or staff members views.

[2]The report, cited as Boruch, Cordray, Pion, and Leviton (1980) in the references, is available from the authors, from the Office of the Assistant Secretary for Management, Evaluation Division, U.S. Department of Education, Washington, D.C., or through the Educational Resources Information Center (ERIC).

[3]The problem of course occurs in popular reporting about statistics as well. See, for example, Kruskal and Mosteller (1979) on "representative sampling."

[4]See Boruch, Anderson, Rindskopf, Amidjiya, and Jansson (1979) for one treatment of the idea, which has probably been broached before, and see Corsi and Hurley (1979) for a remarkable illustration from field experiments on innovation in administrative law.

References

Boruch, R.F., Cordray, D.S., with Pion, G.M. and Leviton, L. An appraisal of educational program evaluations: Federal, state, and local agencies. Report to the Congress. Psychology Department, Northwestern University. Evanston, Illinois June 30, 1980.

Boruch, R.F., McSweeny, A.J., and Soderstrom, E.J. Bibliography: Illustrative randomized field experiments. Evaluation Quarterly, 1978, 4, 655-695.

Boruch, R.F., Anderson, P.S., Rindskopf, D.M., Amidjaya, I.A., and Jansson, D. Randomized experiments for evaluating and planning local programs: A summary on appropriateness and feasibility. Public Administration Review, 1979, 39(1), 36-40.

Conner, R.F. Selecting a control group: An analysis of the randomization process in twelve social reform programs. Evaluation Quarterly, 1977, 1(2), 195-244.

Cook, T.D. and Campbell, D.T. (Eds.) Quasi-experimentation. Design and analysis issues in field settings. Chicago: Rand McNally, 1979.

Corsi, J.R. and Hurley, T.L. Pilot study report on the use of the telphone in administrative fair hearings. Administrative Law Review, 1979, 31(4), 484-524.

Federal Statistical System Project Staff. Improving the Federal Statistical System: Report of the President's Reorganization Project for the Federal Statistical System. Statistical Reporter, May 1980, 197-212.

Kruskal, W. and Mosteller, F. Representative sampling, I: Non-scientific prose. International Statistical Review, 1979, 47, 13-24.

Mosteller, F., Gilbert, S.P. and McPeek, B. Reporting standards and research strategies: Agenda for the editor. Controlled Clinical Trials, 1980, 1(1), 37-58.

Pillemer, D.B. and Light, R.J. Using the results of randomized experiments to construct social programs: Three caveats. In L. Sechrest et al. Evaluation Studies Review Annual, Volume 4, 1979, pp. 717-726 (Beverly Hills, Ca.: Sage).

Riecken, H.W. et al. Social experimentation. New York: Academic, 1974.

Sechrest, L. and Redner, R. Strength and integrity of treatments in evaluation studies. Washington, D.C.: National Criminal Justice Reference Service, National Institute of Law Enforcement and Criminal Justice, Law Enforcement Assistance Administration, U.S. Department of Justice, (June) 1979.