SOCIAL SCIENCE RESEARCH COUNCIL PROJECT ON SOCIAL EXPERIMENTATION

Henry W. Riecken, University of Pennsylvania

The purpose of this paper is to put forward an argument in favor of social experimentation as a method for planning and evaluating social interventions. This general position is the one that has been adopted in a monograph on Social Experimentation that has been prepared by a Committee of the Social Science Research Council and is now in press.

The argument in favor of social experimentation begins from a consideration of the inherent disadvantages of post-hoc "program evaluations". It is clear that program evaluation has recently become a popular activity for social scientists. Much recent social legislation includes a requirement for evaluation of the legislative program. The wave of domestic social reforms in the 1960's that led to compensatory education, community action programs, manpower training, and measures for diminishing racial segregation and sexual discrimination has been responsible for the creation of a mini-industry of evaluation. It is premature to judge how influential such evaluations have been in reshaping social policy, but experience to date suggests there are certain difficulties associated with the usual and ordinary procedure of conducting evaluations of national programs after the fact - that is, waiting until the program is put into full operation before giving appreciable attention to its evaluation. Many of the post-hoc program evaluations that have been carried out on national programs have produced negative evidence or evidence that there has been no change in the state of affairs the program was designed to alter.

This is a persistent difficulty with post-hoc evaluations for a variety of reasons. Many programs, for example, simply do not contain enough variations either in type of treatment; or in range of treatment intensity; or in characteristics of the units affected by the treatment to allow general conclusions to be drawn about alternative treatments. Where programs do contain such variations, treatment effects may be confounded by non-random assignment of the recipients of the treatment. For example, personal characteristics or motivations that lead people to volunteer for a program or to be first in line for a particular treatment may interact with the treatment effects. These are problems of experimental design.

In addition to such difficulties, there are management problems. The very attitude implied by post-hoc evaluations exacerbates certain managerial and institutional frictions. The evaluator turns up after the fact, so to speak, and presumes to make judgments about how well the individuals running the action program have done. The evaluator usually winds up telling them what they should have done if only they had been smarter in the beginning. Such advice often provokes resistance to it on the part of the program operators, who may correctly suspect that the evaluator's hindsight is keener than his foresight. Accordingly one of the management problems of post-hoc evaluation seems to be that the reports written by the consultants who have come in after the event, are received by those who are planning future programs with vast indifference or even hostility.

A final reason for questioning the effectiveness of post-hoc evaluation is that programs, once started, are very difficult to change. Both the clients of the program, and the managers of it get a stake in it. They have a stake in doing things in comfortable and familiar ways and are not very likly to welcome radical innovations, particularly those coming from outside. This is the unreceptivity phenomenon again, though for a different reason.

All four of these considerations suggests that one ought to adopt a somewhat different stance towards social intervention and its evaluation. Instead of establishing a program and then evaluating it, one ought to look at the matter as a cycle of program development, experimental test, and revision prior to installation on a larger scale. The cycle begins with an idea, a notion about how to intervene in a social process, which must then be developed into a program or treatment. To take a concrete case, the idea has been widely circulated in the medical and public health community that a protein-deficient diet of post-weaning children in many less developed countries is responsible for a certain amount of intellectual deficiency at school age and in adulthood. This opinion has resulted in proposals to feed protein supplements from six months of age (or whenever the weaning process is completed) up til school entrance age, when presumably, a substantial amount of physical growth and development has taken place. In order to test this proposition in an experiment it is necessary to develop a feeding program. That is, one must work out a dietary supplement which is acceptable, palatable, and protein-nutritious. One must develop some sort of system for administering the supplement, for making sure that the children who need it get it, that it is not sold in the local market by the families to whom it is given, and that it is not consumed by the adults, but indeed gets to the children who need it. All of these features of a program sound simple once they have been worked out, but they all need to be invented and made a functioning part of the treatment. Incidentally, in the course of developing an experimental treatment from an idea into an operating scheme, a good deal can be learned about potential problems and desirable administrative features of a full-scale program.

Following the development of treatments, the experiment itself must be designed. Since this is an audience of professionals, who know what experimental design is all about, there is no need to go into details about random assignment to treatments, protecting controls from contamination, etc. It should be

emphasized, however, that the SSRC monograph adopts a rather restricted definition of an experiment. The usage in the monograph considers a true experiment as involving at least two treatments - perhaps one active treatment and a control treatment; or two active treatments; together with randomized assignment of treatments to experimental units. This definition is conventional in the statistical literature but quite different from common administrative or bureaucratic usage where "experiment" may mean simply a try-out, a preliminary version of a program that will later be conducted on a larger scale. The usage adopted in the SSRC monograph is much narrower, and conforms to the usual statistical usage.

Two remarks about the design of social experiments may be apropos. One may well ask in respect to variations in the intensity of the treatment whether it might be prudent adding an extremely intensive, even an implausibly intensive treatment, to a design just in order to test whether any treatment at all of the character proposed, at any intensity of application would be effective. Since many social interventions seem weak in comparison with spontaneous counter-forces, it seems worthwhile to inquire whether it is the character of the treatment or merely its intensity of application that produces null or negative results. This is an argument for including treatments that would not be programatically feasible on a national scale since feasibility is of no consequence if even an unfeasibly strong treatment is shown to be ineffective. Secondly, one must face the question of representativeness in the choice of experimental units. It is important to ask: "representativeness for what"? Is the experiment being done for purposes of parameter estimation and generalization to some population? Or is the experimenter simply looking for treatment effects? The answer will determine whether representative sampling of subjects is important.

With the treatment program and the experimental design in hand, an organization must be developed to administer the treatment. For instance, in the food supplementation experiments it is necessary to provide transportation of the raw materials to the experimental site, a place to cook the supplement, a feeding center to which eligible children and mothers come every day, a staff to take care of preparation, serving and clean-up afterward and other matters. This is a rather different task from the technical problems of design or the strategic problems of treatment development and requires different talents. These may not be highly valued by academic institutions, but they are non-negligible. Careless treatment administration can invalidate an experiment just as easily as poor data collection.

Even before the experiment proper is off and running, one must give attention to the analysis and feedback of data for purposes of program revision, policy planning, or perhaps installation of a revised version of the program. This stage brings the cycle of program planning, test, revision and installation to a close.

One might think that, with such a restricted definition of what constitutes an experiment the list of social experiments would be very short. That is, one might suspect that there have been very few randomized, controlled experiments in which the treatments were genuine interventions into societal processes and not merely laboratory exercises. Not so. Robert Boruch's efforts on behalf of the SSCR project have turned up over 120 true randomized social experiments conducted within the last 20 years. This list will be included as an annotated bibliography in the monograph. It covers experiments on such topics as: social rehabilitation programs for juvenile and criminal offenders; law-related programs and procedures; rehabilitation programs in mental health; sociomedical problems and fertility control; assessment of educational and training programs; and many others. The contents of this bibliography demonstrate that randomized experimental tests of social interventions are feasible in a variety of program settings.

Many statisticians are interested in natural approximations to designed social experiments and one section of the SSRC monograph covers these. Donald Campbell, one of the authors of the monograph, has used the phrase "quasi-experiments" to characterize certain natural situations in which something close to an experiment spontaneously and unintentionally occurs. Often, some legislative or administrative change is the virtual equivalent of a deliberately imposed treatment that affects a large segment of a population all at once. One can then observe a time series being interrupted by the change, and look to see whether certain effects have occurred - i.e. whether there have been changes in some dependent variable. An illustration is the administrative decision to require breathalyzer tests of motorists in Great Britain after a certain date, which can be analyzed in relation to rate of car accidents and highway fatalities.

Most of the examples in the SSRC monograph are drawn from the last decade or so of social interventions, but quasi-experiments are, of course, not unique to our own time. I recently discovered a much earlier example that may be of particular interest to this audience because it involves the well known early English bio-statistician, William Farr. In the mid 19th century metropolitan London was the scene of a number of recurrent epidemics of cholera. At that time the disease was not well understood and the mode of its transmission (through water contaminated with fecal matter from infected persons) had not been discovered. Farr had studied cholera epidemics and developed a very

interesting theory about the mode of transmission of the disease*. Farr's study of the data from the 1849 cholera epidemic led him to the conclusion that cholera incidence was related to micro-differences in elevation in various parts of London mediated by differences in miasmas, or atmospheric factors that varied with altitude. His analysis suggested a neat linear progression in incidence varying inversely with 20 foot differences in altitude in the city, leading him to conclude that the lower the altitude the denser the miasmas and hence the higher incidence of cholera. In 1852 the metropolitan government of London passed a law which required that all river water supplied by the private water companies for domestic use must be drawn from the Thames above Teddington Lock, or from tributaries of the Thames above tidal influence. At the time of the 1854 cholera epidemic only one company, the Lambeth Waterworks had complied with the new regulation and had, thereby, suddenly changed its source from one of the most to one of the least contaminated by sewerage. The further important fact is that in a number of districts of the city the Lambeth Company competed directly, street by street, house by house, with the Vauxhall Company which had continued to draw its water supply from a highly polluted portion of the river. In other respects, the social and sanitary conditions of the patrons of the two companies were virtually identical. This situation constituted a vast quasi-experiment involving nearly half a million people for whom the source of drinking water was the most important difference among all social and environmental conditions. By comparing cholera incidence among the customers of the Lambeth Waterworks in the 1854 epidemic with the same customers in the 1849 epidemic; and comparing the experience of patrons of the two companies during the 1854 epidemic, Farr was able to identify the role that drinking water played in the transmission of cholera. (It is of perhaps incidental interest to note that the outcome of this quasi-experiment did not persuade Farr to withdraw his miasmal theory. He simply expanded its scope to include water as well as air among the malignant miasmas).

Quasi-experiments present certain interesting problems of statistical analysis, which arise mainly from the non-random assignment of treatments to units, but that is not the only reason for preferring true randomized experimental design. Appropriate "natural experiments" cannot be counted on to appear in timely fashion to help shape social interventions, programs and policies. Administrators and planners may need to have recourse to randomized experiments for a variety of purposes. One purpose is to estimate parameter values, as in the New Jersey Negative Income Tax Experiment where the problem was to estimate the size of work disincentive effects of a non-conditional income grant. Another purpose is to compare two or more treatments, e.g. the effects on mental development of protein supplementary feeding alone, with the effects obtained from supplying intellectual stimulation along with the protein. A third purpose is to test a concept or claim, as, for example, in the "performance contracting" experiment where certain commercially developed instructional programs in reading and arithmetic were experimentally compared to traditional public school methods to test the claimed superiority of the former.

All of these are, in a sense, specialized and subsidiary purposes that can be encompassed as special cases of the kind of experimental program development, test and revision that was sketched out above and proposed as a substitute for simply installing a program and then evaluating it after the fact. Given the history of social intervention with its abundance of uncertain and sometimes, unintended outcomes, it would seem prudent to learn these lessons on a small, experimental level before going into a nation-wide (or state-wide or company-wide) program that is almost guarenteed to have some flaws, and to be difficult to change or to withdraw.

* For further information about Farr, his theory, and the London cholera epidemics, the reader may consult: Eyler, JM: William Farr on the cholera: The samitarian's disease theory and the statistician's method. Journal of the History of Medicine and Allied Sciences 28:79-100, April, 1973.