# BOOBYTRAPS AND PITFALLS IN THE EVALUATION OF SOCIAL ACTION PROGRAMS

Peter H. Rossi
National Opinion Research Center, University of Chicago

## I: Introduction:

If one were to measure success by the popularity of evaluation research, then empirical social research has certainly arrived. Perhaps, the best example of this popularity lies in the legislation authorizing the present War on Poverty in which the agencies involved are specifically directed to set aside funds for evaluation research. Other ameliorative programs may not give as much formal recognition to such activity, but nevertheless seek social researchers to add to their staffs for this purpose or attempt to get social research centers to provide evaluations of their programs.

There are other measures of success besides popularity. If one were to measure success by the proportion of evaluation researches which are conducted with powerful enough designs to render unequivocal evaluation statements, then empirical social research does not appear to be a smashing success. For a variety of reasons -- some substantive, others related to the present state of development of research methodology, and still others concerned with the "politics" of evaluation -- there are very few evaluation researches which have the elegance of design and clarity of execution which would achieve widespread admiration among social researchers.

The purpose of this paper is to explore some of the main reasons why evaluation research is hard to do well and to suggest some ways in which these difficulties can be overcome. Providing much of the materials on which this paper has been based have been the experiences with such research of the National Opinion Research Center over the past few years. However, I venture that the experiences of other research centers and of individual researchers has not been very different: At least my informal, but undoubtedly highly biased, survey would indicate strong similarities between our experiences and theirs.

In principle, the evaluation of action programs appears to be most appropriately undertaken through the use of experimental designs. All the elements which would strongly recommend such research designs are usually present: The program involved is something which is added to the ongoing social scene by purposive social action as opposed to events which are not under the control of some individual or agency. Because an action program is under someone's control, the construction of experimental and control groups is, in principle, possible. Furthermore, the program is usually not designed to cover an entire population, but only some portion of it so that some of a target population would not be covered, making it possible to think in terms of control groups. Thus, in principle, it is not difficult to design an extremely elegant program of experiments to evaluate the effectiveness of the usual action program. Controlled experiments, however, are not frequently used in evaluation research. For example, there is not a single evaluation research being carried out on the major programs of the War on Poverty which follows closely the model of the controlled experiment.

## II: Action Programs and the Contemporary Scene:

There can be little doubt that the present historical period is one in which there is considerable groping for new and presumably more effective treatments for a variety of presumed ills. We have rediscovered the poor, suddenly become intensely aware that Negroes are an incredibly disadvantaged group, become worried over the plight of the aged, and concerned about a presumed wasteage of brainpower. We also have enough national income to allocate some part of our resources to new programs designed to correct some of the obvious faults in our society.

However, there is an ironic twist to developing a heavy conscience in this historical period. This is because we cannot ordinarily expect that the new treatments we can devise will produce massive results. It appears as if we are in much the same position in the treatment of diseases. The introduction of modern medicine and modern sanitation procedures into a country which has had neither can very dramatically reduce morbidity and mortality, as experiences in some of the emerging nations indicate. But, in the United States of today, each new gain in morbidity and mortality can be expected to be smaller and more difficult to achieve. Providing potable water is much easier to achieve, and more dramatic in its impact on morbidity and mortality, than any attempt we can make to lower the incidence of lung cancer, especially if we try it through lowering levels of smoking in individuals.

Similarly with respect to our social ills. Dramatic effects on illiteracy can be achieved by providing schools and teachers to all children: Achieving a universally high enough level of literacy and knowledge, so that everyone capable of learning can find a good spot in our modern labor force, is a lot more difficult. Hence, the more we have done in the past to lower unemployment rates, to provide social services, etc., the more difficult it is to add to the benefits derived from past programs by

the addition of new ones. Partly, this is because we have acheived so much with the past programs and partly this is because the massive efforts of the past have not dealt with individual motivation as much as with benefits to aggregates of individuals.

In part, the concern of contemporary practitioners in the applied fields with evaluation arises out of their increased methodological sophistication. But, in even larger measure, it arises out of the expectation -- held at some level or other -- that massive effects are not to be expected from new programs and the new treatments are going to be increasingly expensive in terms of time and money. The problem of evaluation in this historical period is that the new treatments can be expected to yield marginal improvements over present treatments and that cost-to-benefit ratios can be expected to rise dramatically. Hence, there is considerable interest in research but considerable apprehension over what it will show concerning the effects of programs.

To illustrate, let us consider the case of Project Headstart: We have apparently wrung most of the benefits we can out of the traditional school system. Although everyone would agree that universal schooling for children up to approximately age sixteen has been a huge success, as opposed to a system of no schooling or of schooling mainly for those to pay for it themselves, there still remains considerable room for improvement, especially in the education of the poor and otherwise disadvantaged. A supplementary pre-school program bringing such children more into parity with those better off because of family background sounds like an excellent program. But, it is hardly likely to produce as much benefit as the introduction of universal elementary schooling did, especially since it is designed to do the job that a full-time institution, the family, neglected to do for one reason or another.

Effective new treatments which produce more than equivocal results are expensive. For example, each trainee at a Job Corps camp costs somewhere between five and ten thousand dollars a year (depending on which estimates you hear), as compared to considerably less than one thousand dollars per year in the usual public high school. Yet a year in a Job Corps Training Center is not going to be five to ten times more effective than a year in a public high school.

Paradoxically, the costs of evaluation are also expensive for these new programs. If effects can be expected to be small, greater precision is needed in research to demonstrate such effects unequivocally. This is another reason why I stressed the controlled experiment as the ideal evaluation research design: Its ability to detect effects is quite powerful compared to alternative methods.

Although as social scientists we can expect the new social programs to show marginal effects, the practitioner does not ordinarily share our pessimism -- at least, not when he faces the Congressional Appropriating Committee. Hence, the claims made in public for the programs are ordinarily pitched much higher, in terms of expectation of benefits, than we could realistically expect with the worst of research and much better than we could expect with the best of research. Thus it turns out that one of the major obstacles to evaluation research is the interests in the maintenance of a program held by its administrators. Their ambivalence is born of a two horned dilemma: On the one hand, research is needed to demonstrate that the program has an effect; on the other hand, research might find that effects are negligible or non-existent.

III: Commitment to Evaluation:

The will to believe that their programs are effective is understandably strong among the practitioners who administer them. After all, they are committing their energies, careers and ideologies to programs of action and it is difficult, under such circumstances, to take a tentative position concerning outcomes. Hence, most evaluation researches which are undertaken at the behest of the administrators of the programs involved are expected to come out with results indicating that the program is effective. As long as the results are positive (or at least not negative) relationships between practitioners and researchers are cordial and sometimes even effusively friendly. But, what happens when it comes out the other way?

A few years ago, the National Opinion Research Center undertook research with the best of sponsorships on the effect of fellowships and scholarships on graduate study in the arts and sciences fields. It was the sincere conviction, on the part of the learned societies which sponsored the research, that such fellowships and scholarships were an immense aid to graduate students in the pursuit of their studies and that heavily supported fields were thereby able to attract better students than fields which were not well supported. The results of the study were quite equivocal: First, it did not appear that financial support had much to do with selection of a field for graduate study. Secondly, it did not appear that graduate students of high quality were being held back from the completion of their graduate programs by the lack of fellowships or scholarships: Those who were committed found some way to get their Ph.D's, often relying on their spouses to make a capital investment in their graduate training. The equivocal nature of the results was quite disappointing to the sponsors whose first reaction was to question the adequacy

of the study's methodology, leading to the coining of a National Opinion Research Center aphorism that the first defense of an outraged sponsor was methodological criticism. The findings affected policy not one whit: The sponsoring groups are still adamantly claiming more and more in the way of financial support for graduate students from the federal government on the grounds that such support materially affects the numbers of talented students who will go to graduate study beyond the B.A., and, furthermore, materially affects the distribution of talent among various fields of study.

Relations between the sponsoring learned societies and our researchers have been cool (if not distant) ever since. The learned societies believe their problem has been badly researched, and the researchers believe that their results have been badly ignored.

Sometimes both the researcher and the practitioner suffer from the will to believe leading to evaluation research containing the most lame sets of qualified results imaginable. Perhaps the best example can be gleaned from the long history of research on the effects of class size on learning. The earliest researches on this topic go back to the beginnings of empirical research in educational psychology and sociology in the early twenties. Since that time there is scarcely a year in which there has not been several dissertations and theses on this topic, not to mention larger researches done by more mature scholars. The researches have used a variety of designs ranging from the controlled experiment to correlational studies, the latest in the series being the results on this score obtained by James Coleman in his nationwide study of schools conducted for the Office of Education under the Civil Rights Act of 1964. The results of these studies are extremely easy to summarize: By and large, class size has no effect on the learning of students, with the possible exception of classes in the language arts. But, the net results of more than two hundred researches on educational ideology and policy has been virtually nil. Every proposal for the betterment of education calls for reductions in the size of classes, despite the fact that there is no evidence that class size affects anything except possibly the job satisfaction of teachers. Even the researchers in presenting their results tend to present them apologetically, indicating the ways in which defects in their research designs may have produced negative findings as artifacts.

In fact, I do not know of any action program that has been put out of business by evaluation research, unless evaluation itself was used as the hatchet to begin with. Why is this the case? Why do negative results have so little impact? The main reason lies in the fact that the practitioners, first of all (and sometimes the researchers), never seriously entertained in advance the possibility that results would come out negative or insignificant. Without committment to the bet, one or both of the gamblers usually welch.

The ways by which welching is accomplished are myriad. It is easy to attack the methodology of any study: Methodological unsophisticates suddenly become experts in sampling, questionnaire construction, experimental design, and statistical analysis, or borrow experts for the occasion. Further replication is called for. But, most often it is discovered that the goals of the program in terms of which it was evaluated are not the "real" goals after all. Thus, the important goals of school systems are not higher scores on multiple choice achievement tests, but better attitudes toward learning, a matter which the researcher neglected to evaluate. Or, the goals of a community organization in an urban renewal area were not really to affect the planning process but to produce a committment to the neighborhood on the part of its residents while the planning took place.

Perhaps the best example of how "real" goals are discovered after goals that were evaluated were found to be poorly attained can be found in the work of a very prominent school administration group. This group, fully committed to the educational modernities of the forties and fifties, found to its surprise that whether or not a school system adopted its programs had little to do with the learning that students achieved. Hence, they dropped achievement tests as a criterion of the goodness of a school or school system and substituted instead a measure of how flexible the administration was in adopting new ideas in curriculum, producing an evaluation instrument which, in effect, states that a school system is good to the extent that it adopts policies that were currently being advocated by the group in question.

IV: Assuring Positive Results:

Given unlimited resources, it is possible to make some sort of dent in almost any problem. Even the most sodden wretch on skid row can be brought to a semblance of respectability for some period of time (provided that he is not too physically deteriorated) by intense, and expensive, handling. But, to make an impact on the denizens of all the skid rows in all of our great cities requires methods that are not intensive and are not expensive case by case. There is not sufficient manpower or resources to lead each single skid row inhabitant back to respectability, if only for a short period.

Yet, many action programs, particularly of the "demonstration" variety, resemble the intensive treatment model. They are bound to produce

results if only because they maximize the operation of the Hawthorne and Rosenthal effects, but cannot be put into large scale operation because either manpower or resources are not available. Hence, programs which work well on the initial run on a small scale with dedicated personnel can be expected to show more positive results than the production runs of such programs with personnel not as committed to the program in question.

The distinction I want to make in this connection is that between "impact" and "coverage." The impact of a technique may be said to be its ability to produce changes in each situation to which it is applied, while the coverage of a technique is its ability to be applied to a large number of cases. Thus, face-to-face persuasion is a technique which has high impact as a means of getting people to come in for physical examinations, but its coverage is relatively slight. In contrast, bus and subway posters may have low impact in the sense of producing a desired effect each time someone is exposed, but large coverage in the sense that many people can be exposed to bus and subway posters very easily.

An extremely effective technique for the amelioration of a social problem is one which has both high impact and high coverage. Perhaps the best example of such techniques can be found in medicine whose immunizing vaccines are inexpensive, easy to administer and very effective in reducing the incidence of certain diseases. It does not seem likely that we will find vaccines, or measures resembling them in impact and coverage, for the ills to which action programs in the social field are directed. It is more likely that we will have action programs which have either high impact or high coverage, but not both. The point I want to emphasize here is that it is a mistake to discard out of hand programs which have low impact but the potentiality of high coverage. Hence, programs which show small positive results on evaluation and which can be generalized to reach large numbers of people can, in the long run, have an extremely significant cumulative effect.

Examples of such programs in the social action field do not easily come to mind. But perhaps an illustration from the field of public health can be cited appropriately: Over the past few decades public health information specialists have been plagued by the fact that their most effective techniques have low coverage and their best mass techniques have little impact. Evaluation research after evaluation research has indicated that it is possible to raise the level of an individual's health knowledge and utilization of health facilities if you can get him to come to a course of lectures on the topic. In contrast, public health information campaigns utilizing the mass media have been shown to have minute effects. Yet, the information of the American population concerning health matters has appreciably increased over the past two decades. It is apparently the case that while no one campaign was particularly effective, their cumulative effects were considerable.

V: The Control Group Problem:

The key feature of the controlled experiment lies in the control exercised by the experimenter over the processes by which subjects are allocated to experimental and control groups. In a well-designed experiment, such allocations are made in an unbiased fashion. But, there are many ways in which a well thought out plan can go awry.

Perhaps the major obstacle to the use of controlled experiments in evaluation research is a political one. The political problem is simply that practitioners are extremely reluctant to allow the experimenters to exercise proper controls over the allocation of clients to experimental and control groups. For example, the proper evaluation of a manpower retraining program requires that potential trainees be separated into experimental and control groups with a contrast being made between the two groups at a later time. This obviously means that some potential clients,who are otherwise qualified, are barred arbitrarily from training - an act which public agencies are extremely reluctant to authorize.

In part, the political problem arises because researchers have not thought through sufficiently the problem of what constitutes a control or non-experimental experience. The logic of experimental design does not require that the experimental group not undergo some sort of treatment, it merely requires that the experimental group not be given the treatment which is being evaluated. In short, we have not been ingenious enough in inventing placebo treatments which are realistic enough to give the public official the feeling that he is not slighting some individuals at random. For example, a placebo treatment for a job retraining program may be conceived of as some treatment designed to help men get jobs but which does not involve retraining and, over which the training program should demonstrate some advantage. Perhaps testing and intensive counseling might be an acceptable placebo for a control group in an experimental evaluation of job training. Or, a placebo treatment for the evaluation of a community mental health center might be referrals to general practitioners for the kinds of treatment they either administer themselves or provide referrals to.

Even in the best circumstances and with the best of sponsors, the carrying out of controlled experiments can run into a number of boobytraps. There is, for

example, the case of an evaluation research all set to go and well designed but whose program did not generate enough volunteers to fill up either the experimental or the control groups. Under these circumstances, the administrator opted to fill up the experimental groups abandoning all attempts at segregating the volunteers into experimental and control groups.

Or, there is the example of a well designed research on the effectiveness of certain means of reaching low income families with birth control information whose design was contaminated by the City Health Department setting up birth control clinics in areas which had been designated as controls!

Or, there is the risk that is run in long range experimental designs that the world may provide experiences to control, which would duplicate in some essential fashion, the experimental treatment. Thus, Wilner et al., in the evaluation of the effects of public housing unfortunately undertook their research in a period when the quality of the general housing stock in Baltimore was being improved at so fast a rate that the contrast in housing conditions between experimental and control groups had greatly diminished by the end of the observational period.

In sum, it is not easy either to obtain sufficient consent to undertake properly controlled experiments or to carry them out when such consent is obtained.

## VI: A Strategy For Evaluation Research:

There are a number of lessons to be drawn from the various sections of this paper which hopefully could go some distance toward devising a strategy for the conduct of evaluation research. While it is true that in a Panglossian best of all possible worlds, the best of all possible research designs can be employed, in a compromised real world, full of evil as it is, it is necessary to make do with what is possible within the limits of time and resources. The problem that faces us then is how can we set up the conditions for doing as best a job we can and produce research which is as relevant as possible to the judgment of the effectiveness of social policy programs.

Although the idea of evaluation research has gained wide acceptance, we are a long way from a full commitment to the outcomes of evaluation research. It is part of the researcher's responsibility to bring to the practitioner's attention that in most cases the effects of action programs are slight and that there is more than an off-chance possibility that evaluation will produce non-positive results. The policy implications of such findings have to be worked out in advance; otherwise the conduct of evaluation research may turn out to be a fatuous exercise.

Secondly, we have a long way to go in devising ways of applying controlled experiments to problems of evaluation. Political obstacles to the use of controls often make it hard to get acceptance of such designs, and the difficulty of maintaining controls in a non-sterile world make full-fledged experimental designs relatively rare in use.

Earlier in this paper, I suggested that we take a lesson from medical research and search for the social analogues of placebos to be administrered to our control groups. There are other directions in which experimental designs should go: For example, considering the high likelihood that treatments have small effects, we need very powerful designs to demonstrate positive results. But because power costs money, it is worthwhile considering research designs which evaluate several types of experimental treatments simultaneously so that the outcomes will be more useful to the setting of program policy. To illustrate: it is considerably more worthwhile to have the results of an experimental evaluation which provides results on several types of Job Corps camps than on job corps camps in general. Looking at the differential effectiveness of several job corps camps provides more detailed and better information for the improvement of job corps programs than would a gross evaluation of the program all told.

This paper has stressed the model of the controlled experiment as the desired one for evaluation research. But, it is abundantly clear that for a variety of reasons, controlled experiments are rarely employed as evaluational devices and that they are difficult to employ. Most frequent are some sort of quasi-experiments in which the control groups are constructed by methods which allow some biases to operate and correlational designs in which persons subjected to some sort of treatment are contrasted with persons who have not been treated, controlling statistically for relevant characteristics.

The important question which faces the evaluation researchers is how bad are such "soft" evaluational techniques, particularly correlational designs? Under what circumstances can they be employed with some confidence in their outcomes?

First of all, it seems to me that when it is massive effects that are expected and desired, "soft" techniques are almost as good as subtle and precise ones. To illustrate, if what is desired as the outcome of a particular treatment is complete remission of all symptoms in each and every individual subject to treatment, then it is hardly

necessary to have a control group. Thus if a birth control technique is to be judged effective if and only if it completely eliminates the chance of conception in an experimental group, then the research design is vastly simplified. The question is not whether those who use the method have less children than those who do not, but whether they have any children at all, a question which can be easily decided by administering the technique to a group and counting births (or conceptions) thereafter.

The obverse of the above also holds. If a treatment which is to be tested shows no effects using a soft method of evaluation, then it is highly unlikely that a very precise method of evaluation is going to show more than very slight effects. The existence of complex and large interaction effects which suppress large differences between a group subject to a treatment and statistical control groups seems highly unlikely. Thus if children participating in a Head Start program show no gain in learning ability compared to those who did not participate in the program, holding initial level of learning constant, then it is not likely that a controlled experiment in which children are randomly assigned to experimental and controlled groups is going to show dramatic effects from Head Start programs.

Of course, if a correlational design does show some program effects, then it is never clear whether selection biases or the program itself produce the effects shown.

This means that it is worthwhile to consider soft methods as the first stage in evaluation research, discarding treatments which show no effects and retaining those with opposite characteristics to be tested with more powerful designs of the controlled experimental kind.

Although ex post facto designs of a correlational variety have obvious holes in them through which may creep the most insidious of biases, such designs are extremely useful in the investigation of effects which are postulated to be the results of long acting treatments. Despite the fact that it is possible that cigarettes cause cancer, the evidence from ex post facto studies of the correlation between cigarette smoking and lung cancer can hardly be ignored, even though the evidence is not pure from the viewpoint of a purist. Similarly, NORC's study of the effects of Catholic education on adults, despite all our efforts to hold constant relevant factors, can still be easily produced by self selection biases that were too subtle for our blunt instruments to detect. We have nevertheless gained a great deal of knowledge concerning the order of effects that can be expected, were a controlled experiment extending over a generation conducted. The net differences between parochial school Catholics and public school Catholics are so slight that we now know that this institution is not very effective as a device for maintaining religiosity and that furthermore the effects we found are quite likely to have been generated by selection biases.

From these considerations a strategy for evaluation research is beginning to emerge. It seems to me to be useful to consider evaluation research in two stages--a Reconnaissance Phase in which the soft correlational designs are used to screen out those programs it is worthwhile to investigate further; and an Experimental Phase in which powerful controlled experimental designs are used to evaluate the differential effectiveness of a variety of programs which showed up as having sizable effects in the first phase.