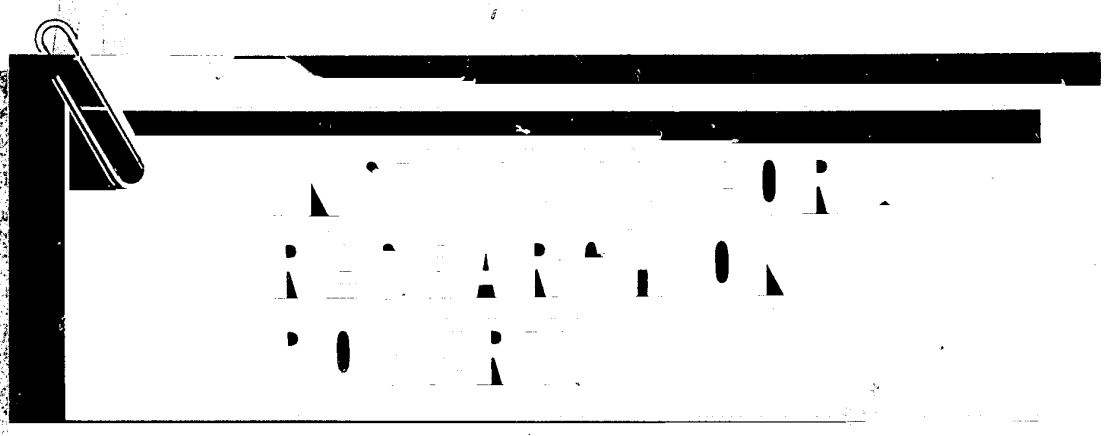
42-69



THE METHODOLOGY OF EVALUATING SOCIAL ACTION PROGRAMS

by

GLEN G. CAIN

ROBINSON G. HOLLISTER

THE UNIVERSITY OF WISCONSIN-MADISON, MADISON, WISCONSIN

3151 75

4

THE METHODOLOGY OF EVALUATING SOCIAL ACTION PROGRAMS

Glen G. Cain

Robinson G. Hollister

This research was supported by funds granted to the Institute for Research on Poverty, University of Wisconsin, pursuant to the provisions of the Economic Opportunity Act of 1964. Professor Cain and Professor Hollister are associated with the University of Wisconsin Department of Economics and are members of the Institute staff. The authors are grateful to the following persons, who have increased their understanding of the ideas in this paper or have commented directly on an earlier draft (or have done both): David Bradford, Frank Cassels, John Evans, Woodrow Ginsberg, Thomas Glennan, Robert Levine, Guy Orcutt, Gerald Somers, Ernst Stromsdorfer, Harold Watts, Arnold Weber, Burton Weisbrod, and Walter Williams. Shorter versions of this paper are scheduled to appear in the 1969 annual research volume of the Industrial Relations Research Association, *Public and Private Manpower Policies*, and the volume consisting of the Proceedings of the North American Conference on Cost-Benefit Analyses, held in Madison, Wisconsin, May 14-15, 1969.

APOLOGIA

This paper is largely motivated by our experiences as academics who became directly enmeshed in the problems of a public agency which was under considerable pressure--generated by both the agency staff itself and external factors--to "evaluate" manpower, and other social action, programs.

It became evident that there were several major obstacles to effective evaluation in this context. These obstacles were created both by the several types of "actors" necessarily involved in such evaluation efforts and by complications and weaknesses in the theory and methodology to be applied. Difficulties of communication among the "actors", due both to differences in training and to suspicions about motives, often made it hard to distinguish between difficulties arising because the theory was weak and those arising because adequate theory was poorly understood.

In this paper we try to separate out some of these issues, both those concerning the adequacy of theory and methodology and those relating to the various sorts of actors. We have sought to couch the discussion in language that will make it available to academics, who we feel need a heightened awareness of the more practical difficulties of execution of evaluations in the social action context--and to public agency and political personnel, who we believe would benefit from increased sensitivity to the ways in which careful consideration of the design and careful control of evaluations can increase the power of the information derived from such efforts. The attempt to reach both audiences in one paper produces a mixture of elements bound to strike members of either audience as, at some points, extremely naive and, at others, disturbingly recondite. We can only hope that such reactions will be transformed into a resolve to initiate a more meaningful dialogue on these issues, a dialogue we feel is crucial to the development of an effective approach to evaluations of social action programs.

TABLE OF CONTENTS

THE METHODOLOGY OF EVALUATING SOCIAL ACTION PROGRAMS

Glen G. Cain and Robinson G. Hollister

Manpower programs used to consist almost entirely of vocational training and various but limited types of assistance for the worker searching for jobs within local labor markets. But with the recent emphasis on problems of poverty and the disadvantaged worker, manpower programs have come to involve remedial and general education, to intermesh with community action programs providing a variety of welfare services, and, on a trial basis, to assist in migration between labor markets. They are part of a broader class of programs which, for lack of a better term, we might call social action programs. Our paper will include many references to this broader class, and in particular to anti-poverty programs. In so doing, we hope to provide a more general and more relevant perspective on the topic of evaluation methodology.

We hold the opinion, apparently widely shared, that existing evaluations of social action programs, (and we are including our own), have fallen short of meeting the standards possible within the disciplines of the social sciences. The reasons for these shortcomings are easy to identify. The programs typically involve investments in human beings, a relatively new area of empirical research in economics. They are aimed at such social and political goals as equality and election victories, as well as economic objectives concerning, say, income and employment. They often attempt to deliver services on a large enough scale to make a noticeable impact upon the community. And at the same time, they

are expected to provide a quasi-experimental basis for determining what programs ought to be implemented and how they ought to be run.

It is not surprising then, that evaluations of social action programs have often not been attempted and when attempted, have not been successful. Despite this background, we believe that existing data and methods permit evaluations which, while not satisfying the methodological purists, can at least provide the rules of evidence for judging the degree to which programs have succeeded or failed. Specifically, the theme we will develop is that evaluations should be set up to provide the ingredients of an experimental situation: a model suitable for statistical testing, a wide range in the values of the variables representing the program inputs, and the judicious use of control groups.

The paper reflects several backgrounds in which we have had some experience--from economics, the tradition of benefit-cost analyses; from the other social sciences, the approach of quasi-experimental research; and from a governmental agency, the perspective of one initiating and using evaluation studies. Each of these points of view has its own literature which we have by no means covered, but to which we are indebted.[1]

TYPES OF EVALUATION

There are two broad types of evaluation. The first, which we call "process evaluation," is mainly administrative monitoring. Any program must be monitored (or evaluated) regarding the integrity of its financial transactions and accounting system. There is also an obvious need to check on other managerial functions, including whether

or not accurate records are being kept. A component of process evaluations are progress reports aimed at determining the need for possible administrative changes in the operation of the program. In sum, "process evaluation" addresses the question: Given the existence of the program, is it being run honestly and administered efficiently?

A second type of evaluation, and the one with which we are concerned, may be called "outcome evaluation," more familiarly known as "cost-benefit analysis." Although both the inputs and outcomes of the program require measurements, the toughest problem is deciding on and measuring the outcomes. With this type of evaluation the whole concept of the program is brought into question, and it is certainly possible that a project might be judged to be a success or a failure irrespective of how well it was being administered.

A useful categorization of cost-benefit evaluations draws a distinction between a priori analyses and ex post analyses. An example of a priori analysis is the cost-effectiveness studies of weapons systems conducted by the Defense Department, which have analyzed war situations where there were no "real outcomes" and, thus, no ex post results with which to test the evaluation models. Similarly, most evaluations of water resource projects are confined to alternative proposals where the benefits and costs are estimated prior to the actual undertaking of the projects.[2] Only in the area of "social action" programs such as poverty, labor training, and to some extent housing, have substantial attempts been made to

evaluate programs, not just in terms of before-the-fact estimates of probable outcomes or in terms of simulated hypothetical outcomes, but also on the basis of data actually gathered during or after the operation of the program.

A priori cost-benefit analyses of social action programs can, of course, be useful in program planning and feasibility studies, but the real demand and challenge lies in ex post evaluations. This more stringent demand made of social action programs may say something about the degree of skepticism and lack of sympathy Congress (or "society") has concerning these programs, but this posture appears to be one of the facts of political life.

Two additional differences between human investment programs and physical investment programs deserve mention--although whether these differences are real or merely apparent is a debatable point. One is the complexity of behavioral relations which the social action programs try to change. Is it correct to say that these relations are more difficult to analyze and predict than the technological relations which appear in defense and water resource analysis? Perhaps, but if the analysis of the latter really requires data on propensities of aggressive behavior or on values of recreational activities, respectively, then we may question whether these are easier to analyze than, say, employment behavior. A second difference is the shorter history and subsequent dearth of analytic studies of social action programs, a fact clearly related to the weaknesses of our theory and empirical knowledge of the behavioral relationships affected by the policies.

An awareness of these rather basic differences between the evaluations (or benefit-cost analyses) which have been carried out allegedly with some speed and success in other areas and evaluations which have been looked for and generally not been forthcoming in the social action area is important in understanding the relatively "poor performance" of evaluators in the latter area. We can then be better prepared to recognize that the methodology for evaluation of social action programs will have to be developed in new ways to cope with their special difficulties.

PROBLEMS OF THE DESIGN OF THE EVALUATION

A. Specification of the Objectives

In the methodology of program evaluation which has been constructed, one of the principal tenets is that the first step in the analysis must be to specify the objectives of the program. Unfortunately, agreement on this principle has not facilitated its implementation, the problem being that few programs have a single clearly defined objective or even one dominant objective.

It becomes necessary to assign weights to the different objectives and to guard against both double-counting and undercounting. Arguments arise concerning "ultimate" objectives and "intermediate" objectives, and there will usually be a struggle to agree upon some measurable intermediate objectives which can serve as proxies for (practically speaking) unmeasurable ultimate objectives. Economists, who deal theoretically with the concepts

of "welfare" and "utility" while their empirical work involves incomes and prices, should not find it difficult to appreciate the legitimacy of non-measurable entities.

We suggest, however, that in general the measures of program outputs, which may be proxies for ultimate objectives, should be measures of behavior and of tangible changes, such as income change, employment gain, and educational attainment. Lower priority should be given to the less tangible measures of self-images, community images, and opinion polls of peoples' attitudes towards the programs. The defense of this position rests mainly on the practical grounds of choosing outcomes which may be more accurately measured, both immediately and in terms of measures of outcomes, and choosing those which are more stable as predictors of a longer run or permanent assessment. We would argue for example that the relatively hard measures of cognitive educational gain are a more reliable and valid measure of the benefits of a Head Start program than are surveys of parents' or teachers' attitudes about the program. The latter should not be ignored, only given less weight. We suggest that, over the long run *but not necessarily in the short run*, attitudes will closely correlate with the more tangible performance indicators. So, why not aim right from the beginning at measuring the program's substance rather than its public relations effects?

Although some measurable objectives are necessary for all but the crudest, journalistic type of evaluation, not all such objectives provide an obvious or easy translation into dollars to permit the

desired benefit-cost calculation. In our judgment and experience, however, the problem of assigning dollar values is a step we seldom reach because we are unable to measure in the first instance the more direct or specific program outcome. Our failures in this respect are numerous--witness Head Start, health programs, and many of the manpower programs in which we simply do not know what difference the program has made. It is absolutely necessary that we first concentrate on assessing the change in educational attainment, in health, in employment and earnings or in whatever the program objective is. If this is done, we as economists may then offer some guides regarding the dollar worth of these changes, but even if the policy-maker decides on his own system of pricing, we will have constrained the possibilities for mistaken judgments.

Indeed, the problems of specifying objectives will not disappear even if there is agreement on a translation of program outcomes to dollar values. Consider a program which provides for a simple transfer of money to the participant, who, let us assume, is poor. Obviously, the objective of improving the economic status of the participant is unambiguously attained, but are we satisfied with this objective? It is instructive to begin any discussion of the objectives of social action programs aimed at the poor or disadvantaged person with a simple income-transfer program, because all the arguments about self-help, non-economic goals, and community-wide goals can be explicitly aired. Economists in particular are forced to face these issues and will be better prepared for them when they arise, sometimes in

disguised forms in analyses of more complicated programs of assistance. At the same time, when non-economists are directly confronted with the example of a simple income transfer program, they will be able to better understand and accept the extent to which such a transfer program is the implicit criterion of a benefit-cost ratio of 1, as used in benefit-cost analysis.

Specifying program objectives is an important step, but there is a risk that the attempt to reach unanimous agreement on the whole hierarchy of intermediate and ultimate objectives will become a road-block to the undertaking of program evaluations. There have been numerous cases in which months, and even years, have been taken up in arguments over what the program objectives "really are" or how multiple objectives are to be "weighted" to add up to some over-all measure. In the meantime, programs have stumbled on with no evaluation or new programs have been forestalled because no a priori evaluation was undertaken to assess the feasibility of the program. Wiley bureaucrats have been able to prevent evaluation of their programs for many months by refusing to "sign off" on a defined set of objectives. (The legislative history of a program, like the Scriptures, provides a boundless source of Pharisaical counter-interpretations as to intended objectives).

In the same vein, it must be recognized that there are some important social action programs for which it is necessary to observe what a program is doing and, in the process of observation, identify

what the objectives are. Some programs leave considerable operational discretion to the local level, so that the program as actually implemented may differ considerably from area to area. In others, the legislative or administrative mandate may reflect a compromised mixture of several loosely related program proposals.[3]

In these types of programs what is necessary is something which might be called a "search-evaluation," and attempts to follow the usual dogma of evaluation, starting with the definition of a single objective--or a hierarchy of objectives--for the program, are bound to fail. The first stages of the evaluation must be to find the actual nature of the program in various areas. Of course, some sort of theory is required to suggest which objectives are relevant, but the search process may modify our theory. An iterative procedure is called for in which the process of evaluation goes on simultaneously with a search for the objectives of various elements of the program.

An obvious example of the type of program which requires a search evaluation is the Community Action Program. It embodies both a legislative compromise of quite different proposals and considerable latitude for local discretion in implementation.[4] Early attempts to initiate an evaluation of the program, both overall and for its components, foundered on conflicts over the definition of objectives of the program. Participation of the poor, institutional change, more efficient delivery of services, and mobilization and coordination of existing federal, state and local resources were among those advocated as primary objectives. Evaluation of the

program only began to move forward when a strategy of evaluation was adopted which had an initial search phase.

It should be clear that search evaluation situations--with the Community Action Program as an example--reflect in an extreme form most of the problems outlined above. It is almost tautological to note that it is the *ex post* nature of the evaluation that necessitates the "search" phase. The problem of difficult-to-measure objectives is also related, since part of the evaluation process consists of a search for adequate measures of what have heretofore been regarded as qualitative phenomena. (How does one quantify institutional change?) Finally, these problems are related to the poorly conceptualized behavioral content in such program elements as "participation" and "institutional change".

It may be helpful, in sum, to suggest that the structure of the dogma of evaluation developed in defense and water resources was largely a deductive structure, whereas the structure suggested for "search evaluation" situations is essentially, in its initial phases, inductive in nature. Analysts familiar with the first type are reluctant to accept the latter. In certain situations, however, the choice is between a "search evaluation" or no evaluation.

B. The Use of Control Groups

Given the objective of the program, the question, "What difference did the program make?", should be taken literally. We want to know the difference between the behavior with the program and

the behavior if there had been no program. To answer the question, some form of control group is essential. We need a basis for comparison --some base group that performs the methodological function of a control group. Let us consider some alternatives.

The Before-and-After Study. In the before and after study, the assumption is that each subject is his own control (or the aggregate is its own control) and that the behavior of the group before the program is a measure of performance that would have occurred if there had been no program. However, it is well known that there are many situations in which this assumption is not tenable. We might briefly cite some examples found in manpower programs.

Sometimes the "before situation" is a point in time when the participants are at a particularly low state--lower, that is, than is normal for the group. The very fact of being eligible for participation in a poverty program may reflect transitory conditions. Under such conditions we should expect a "natural" regression toward their mean level of performance if we measure their status in an "after situation," even if there were no program in the intervening period. Using zero earnings as the permanent measure of earnings of an unemployed person is an example of attributing normality to a transitory status.

Another similar situation is when young people are involved, and the "natural" tendency over the passage of time would be expected to be improvement in their wages and employment situation.

There may be some structural change in the personal situations of the participants before and after the program, which has nothing to do with the program but would vitiate any simple before-and-after comparison. We should not, for example, look upon the relatively high earnings record of coal miners or packinghouse workers as characteristic of their "before situation" if, in fact, they have been permanently displaced from their jobs.

As a final example of a situation in which the before-and-after comparison is invalid, there is the frequent occurrence of significant environmental changes--particularly in labor market environments--which are characterized by seasonal and cyclical fluctuations. Is it the program or the changed environment which has brought about the change in behavior? All of the above examples of invalidated evaluations could have been at least partially corrected if the control groups had been other similar persons who were in similar situations in the pre-training period.

Control Groups Which are not Program Participants: Small Group Studies Versus Large Group Studies. The particular strength of the small scale study is that it greatly facilitates the desideratum of random assignments to "treatment groups" and "control groups" or, at least, a closely supervised matching of treatment and control groups. Its particular shortcoming is that it is likely to lack representativeness--both in terms of the characteristics of the program participants and in terms of the character of the program. There is first the problem of a "hot house environment" of the small group

study. (See discussion of "replicability" below.) Second, a wide range of values of the program inputs (i.e., in terms of levels of a given treatment or in terms of qualitatively different types of treatments) is less likely to be available in a small group study. (See the discussion on "statistical considerations" below). Third, the small group study may not be able to detect the program's differential effects on different types of participants (e.g., by age, sex, color, residence, etc.,) either because the wide variety of participant types are not available or because their numbers are too small. Finally, it is both a strength and a weakness of the small scale study that it is usually confined to a single geographic location. Thus, although "extraneous" noise from different environment is eliminated, we may learn little or nothing about how the program would operate in different environments.

The large scale study, which involves gathering data over a wide range of environments, customarily achieves "control" over the characteristics of participants and nonparticipants and over programs and environmental characteristics by statistical methods, rather than by randomization or careful matching, individual by individual. These studies have the capability of correcting each of the shortcomings attributed to the small scale studies in the preceding paragraph. But because they are almost impossible to operate with randomization, the large scale studies run afoul of the familiar problem in which the selectivity of the participants may

be associated with some unmeasured variable(s) which makes it impossible to determine what the net effect of the treatment is. Since this shortcoming is so serious in the minds of many analysts, particularly statisticians, and because the small scale studies have a longer history of usage and acceptability in sociology and psychology, it may be worthwhile to defend at greater length the large scale studies, which are more common to economists.

Randomization is seldom attempted for reasons having to do with the attitudes of the administrators of a program, local pressures from the client population, or various logistic problems. Indeed, all these reasons may serve to botch an *attempted* randomization procedure. Furthermore, we can say with greater certitude that the ideal "double-blind experiment with placebos" is almost impossible to achieve. If we are to do something other than abandon evaluation efforts in the face of these obstacles to randomization, we will have to turn to the large scale study and the statistical design issues that go along with it.

The fact that the programs vary across cities or among administrators may be turned to our advantage by viewing these as "natural experiments"[5] which may permit an extrapolation of the results of the treatment to the "zero" or "no-treatment" level. This latter device may be particularly useful if the analyst can work with the administrator in advance to design the program variability in ways which minimize the confounding of results with environmental influences. Furthermore,

ethical problems raised by deliberately excluding some persons from the presumed beneficial treatments are to some extent avoided by assignments to differing treatments (although, here again, randomization is the ideal way to make these assignments).

It is difficult, at this stage, to provide more than superficial observations regarding the choice between small and large-scale studies. It would seem that for those evaluations that have a design concept which is radically different from existing designs or where there is a quite narrow hypothesis which requires detailed examination, a small group study would be preferable. Conversely, when the concept underlying a program is quite broad and where large amounts of resources are to be allocated, the large group approach is probably more relevant--a point argued in greater detail in our discussion of the "replicability criterion."

C. The Replicability Criterion

A source of friction between administrators of programs and those doing evaluation research, usually academicians, is the failure to agree upon the level of decision-making for which the results of the evaluation are to be used. This failure, which is all the more serious because the issue is often not explicitly addressed, leads to disputes regarding two related issues--the scope of the evaluation study and the selection of variables to be studied. To deal with these disputes, we suggest applying the "replicability criterion." We apply this name to the criterion because of the large number of cases in which evaluations of concepts have been made on the

basis of projects which are not likely to be replicable on a large scale or which focus on characteristics of the project which are not within the ability of decision-makers to control. To take an extreme example, it has sometimes been stated that the success of a compensatory education program depended upon the "warmth and enthusiasm" of the teachers. In the context of a nationwide program, no administrator has control over the level of "warmth and enthusiasm" of teachers.

It is sometimes argued by administrators that evaluations which are based upon samples drawn from many centers of a program are not legitimate tests of the program concept since they do not adequately take into account the differences in the details of individual projects or of differentiated populations. These attitudes frequently lead the administrators or other champions of the program to select, either ex ante or ex post, particular "pet" projects for evaluations that "really count." In the extreme, this approach consists of looking at the successful programs (based on observations of ongoing or even completed programs) and then claiming that these are really the ones that should be the basis for the evaluation of the program as a whole. *If* these successful programs have worked with representative participants in representative surroundings and *if* the techniques used--including the quality of the administrative and operational personnel--can be replicated on a nationwide basis, *then* it makes sense to say that the evaluation of the particular program can stand for an evaluation of the overall program. But we can seldom assume these conditional statements. After all, each of the individual programs, a few political plums notwithstanding, was set up because someone thought it was worthwhile. Of

course, some will flop because of poor teachers or because one or more operations were fouled up--but it is in the nature of the beast that some incompetent administrative and operational foul-ups will occur. A strength of summary, over-all measures of performance is that they will include the "accidental" foul-ups with the "accidental" successes, the few bad administrators and teachers as well as the few charismatic leaders. As a case in point, consider the success (according to prevailing opinion) of Reverend Sullivan's Operation Industrial Council in Philadelphia with the (as yet) absence of any evidence that the OIC idea has been successfully transferred elsewhere.[6]

Small scale studies of pre-selected particular programs are most useful either for assessing radically different program ideas or for providing the administrator with information relevant to decisions of program content *within* the confines of his overall program. These are important uses, but the decisions at a broader level which concern the allocation of resources among programs of widely differing concepts call for a different type of evaluation with a focus on different variables.

It may be helpful to cite an example of the way in which the replicability criterion should have been applied. A few years ago, a broad scale evaluation of the Work Experience Program[7] was carried out. (The evaluation was of necessity based upon very fragmentary data, but we are here concerned with the issues it raised rather than with its own merits.) The evaluation indicated that on the average

the unemployment rates among the completers of the program were just as high as those with similar characteristics who had not been in the program. On the basis of this evaluation, it was argued that the concept of the program was faulty, and some rather major shifts in the design and in the allocation of resources to the program were advocated.[8] Other analysts objected to this rather drastic conclusion and argued that the "proper" evaluative procedure was to examine individual projects within the program, pick out those projects which had higher "success rates," and then attempt to determine which characteristics of these projects were related to those "success rates."[9]

The argument as to which approach is proper depends on the particular decision framework to which the evaluation results were to be applied. To the administrators of the program, it is really the project by project type of analysis which is relevant to the decision variables which they control. The broader type of evaluation would be of interest, but their primary concern is to adjust the mix of program elements to obtain the best results within the given broad concept of the program. Even for program administrators, however, there will be elements and personnel peculiar to a given area or project that will not be replicable in other areas and other projects.

For decision-makers at levels higher than the program administrator the broader type of evaluation will provide the sort of information relevant to their decision frame. Their task is to allocate resources among programs based upon different broad concepts. Negative findings

from the broader evaluation argue against increasing the allocation to the program, although a conservative response might be to hold the line on the program while awaiting the more detailed project-by-project evaluation to determine whether there is something salvagable in the concept embodied in the program. There will always be alternative programs serving the same population however, and the decision-maker is justified in shifting resources toward those programs which hold out the promise of better results.

The basic point is that project-by-project evaluations are bound to turn up some "successful" project somewhere, but unless there is good evidence that that "success" can be broadly replicated and that the administrative controls are adequate to insure such replication, then the individual project success is irrelevant. Resources must be allocated in light of evidence that concepts are not only "successful" on a priori grounds or in particular small-scale contexts but that they are in fact " successful" in large-scale implementation.

D. The Theoretical Framework--Some Statistical Considerations.

The main function of a theoretical framework in cost-benefit evaluations is to provide a statistical model suitable for testing. A discussion of the economic content of the statistical model is taken up in the next section; here we focus on more general questions of the statistical design of the evaluation. Generally, it makes little or no difference whether the statistical method is analysis of variance, regression analysis, or simply working with cell values in

tables, but we will adopt the terminology of the regression model for purposes of this discussion. In this model, the dependent variable is the objective of the social action program and the particular set of independent variables of most interest to us are those that describe or represent the program, or program inputs. In this discussion the independent variables will sometimes be referred to as "treatment variables."

Usually our theory (which includes the body of substantive findings from previous studies) can tell us *something* about what variability can be expected in the behavior described by the dependent variable, and this information is necessary for determining the appropriate sample size. On the same issue, the theory can tell us what independent variables may be included as statistical controls for the purpose of reducing the unexplained or residual variation in the dependent variable. Clearly, the smaller the residual variation is, the smaller is the sample size needed to attain a given level of precision (or statistical significance) in our results. Another way of making this point is to say that the smaller the residual variation the greater is the statistical significance we achieve for a given sample size.

As an example of these considerations, assume that the objective of the program is to improve the wage earnings of a group of low-wage workers. Our dependent variable is some measure of earnings over a period of at least one year after those who were in the training program had left it. We can say at the outset that on the basis of

the existing studies of income variability, we should be prepared for a large variation in the earnings of our subjects--standard deviations in the hundreds of dollars would be typical. Moreover, these same studies combined with other a priori information can indicate what independent variables (like the worker's age, education, etc.) will account for some of this variation and thereby produce a smaller residual variation. We might add that the existing studies of determinants of earnings indicate that we should expect a relatively large residual variation to remain. Thus, we might still have to contend with unexplained variability (or standard errors of estimates) in the hundreds of dollars per subject.

How serious is a large residual variation in terms of preventing the detection of an effect of some training program? This depends on how large an effect we expect the training program to bring about, or, in more technical terms, it depends on the size of the partial regression coefficients representing the programs. Here again, our existing theory can narrow the range of our ignorance. Thus, we might be able to combine our information on the amount of variability in the dependent variable, earnings, with educated guesses about the earnings effect of a training program to permit us to decide how large a sample will be required to achieve some selected confidence interval on our estimates.[10] Suppose that we have, for example, relevant studies of the effects of investments in education or training suggesting that rates of return of 5 to 25 percent might be expected. Thus, on an investment of $1,000 the annual earnings of a worker

might be raised by $50 to $250.[11] Obviously, for the given level of significance, a large sample will be required and/or more statistical controls will be necessary to detect changes of this order of magnitude than if the program were expected to increase earnings of the participant by $1000.

Indeed, it is precisely programs which have large and dramatic effects which can be evaluated with a loose design and an almost journalistic level of evaluation, but we would contend that almost all social action programs, and particularly those in the field of manpower training and education, are unlikely to bring about such spectacular changes.[12] Regarding the *results* of a program, the analogy between a Salk vaccine for polio and a social action treatment for poverty does not hold. The irony is that regarding the *means* of evaluation, in many ways the test of the Salk vaccine provides an excellent model for social scientists to study.

Up to now we have discussed the role of theory in providing information on expected variability in the dependent variable representing the goals of the program and on the expected effect of various independent variables--effects of treatment representing the program and of control variables which help reduce the residual variation in the dependent variable. Note that the failure to attain statistical significance of the effect of the treatment variable because of either a large unexplained variation in the dependent variable or small effects of treatment variables, can be overcome with sufficiently large sample sizes. But in our opinion, the most serious defect in evaluation studies are biases in the measures of

effects of the treatment variables, and this error is unlikely to be removed by enlarging the sample size.

One source of bias is inaccurate measures of the treatment variable, but a more pervasive and more serious problem is the presence of variables, not included in the statistical model, which are correlated with both the dependent variable and the treatment variable. Had the assignment to a program been made on a random basis, the laws of probability would have assured a low correlation (zero in the limit of a large enough sample size) between participation in the program and these omitted variables. In the absence of randomization, we must fall back on statistical controls. At this point our theory and a priori information are crucially important. The requirements are obvious: to identify the variables whose omission leads to biases in the measured effects of the treatment variables and to include them in the model. These variables may be objectively measurable, such as age or education or previous work experience. Or they may be such difficult-to-measure characteristics as ambition, motivation, or an "appealing personality."[13]

As we know too well, however, our theories are woefully weak in providing us with the correct list of variables for explaining such dependent variables as income change, employment experience, health, status, or educational attainment, and we often do not have measures of those we do know about. The latter problem frequently arises because of the unfortunate practice of inviting the evaluator in *after* the program has been run and the data have been collected.

Even in the best of situations regarding the availability of objective measures of important variables, if we do not have random assignments we must still admit the possibility that *self-selectivity* or the *selectivity procedures* of the program administrators has introduced a systematic difference between the participants and the nonparticipants. We do not claim, as the purists would, that non-random procedures invalidate all evaluations, although there are cases when they undoubtedly have, but the advantages of randomization are immense and we can do a great deal more to achieve this procedure if we can only convince each other of its importance. It is clear that those responsible for the tests of the Salk vaccine were convinced.

Another important advantage of randomization should be mentioned. We have noted that variables which are correlated with both the treatment variable and the dependent variable must be included in the model to measure treatment effects without bias. However, since our information about the effect of the treatment variable necessarily depends on variability in treatments, and since the only variation we can observe within the framework of the statistical model is the residual variation in treatments--that is, variation which remains after the entire set of independent variables is included, greater efficiency is obtained when the treatment variable is uncorrelated with the other independent variables. In the opposite extreme, if the treatment variables were perfectly correlated with some other variable or combination of variables, we would be unable to distinguish between which of the two sets of factors caused a change. It follows that even in the absence of randomization, designing the

programs to be studied with as wide a range in levels and types of "treatments" as possible will serve to maximize the information we can extract from an ex post analysis.

There are reasons in addition to those of statistical efficiency for planning for a wide range of values in the treatment of programmatic variables. One is that social action programs have a tendency to change, rather frequently and radically, during the course of their operation. Evaluations designed to test a single type of program are rendered meaningless because the program-type perishes. But of the design covers a wider variety of programs, then a built-in hedge against the effects of change is attained. Indeed, there is an even more fundamental reason why a wide range of inputs and program types should be planned for, and it is simply this: we seldom know enough about what will work in a social action program to justify putting our eggs in the single basket of one type of program. This evaluation model for a single type of project, sometimes described as the analogue of the "pilot plant," is not the appropriate model for social action programs given our current state of knowledge.[14]

E. The Theoretical Framework--Some Economic Considerations.

For operational purposes we will assume that the evaluation of each social action program can, at least in principle, be cast in the statistical model discussed in the previous section, complete with variables representing an objective of the program, treatment variables representing the program inputs, control variables, and control groups.[15] However, the substantive theoretical content of these models--the particular selection of variables and their

functional form--must come from one or more of the traditional disciplines such as educational psychology (e.g., for Head Start), demography (e.g., for a family planning program), medical science (e.g., for a neighborhood health center), economics (e.g., for a manpower training program), and so on.

Sooner or later economics must enter all evaluations, since "costing out" the programs and the setting of implicit or explicit dollar measures of the worth of a program are essential steps in a complete evaluation. And this is true even though the most difficult part of the evaluation may lie in determining what the specific program effects are in terms of educational achievement, health, or some other nonmonetary benefit.

In making the required cost-benefit analysis, the part of economic theory that applies is the investment theory of public finance economics, with its infusion of welfare economics. The function of investment theory is to make commensurable inputs and outcomes of a social action program which are spaced over time. Welfare economics analyzes the distinctions between financial costs and real resource costs, between direct effects of a program and externalities, and between efficiency criteria and equity (or distributional) criteria.

We will say very little on the last mentioned distributional or equity question of *who pays* and *who receives*, even though we strongly feel that accurate data on the distribution of benefits and costs is essential to an evaluation of social action programs. However, the task of conducting a "conventional" benefit-cost analysis (where the criterion is allocative efficiency) is sufficiently complex that we believe it preferable to separate the distributional questions.

Program Inputs. In the investment theory model costs are attached to all inputs of a program and a single number emerges which measures the present value of the resources used. Although the purpose of this procedure is to reduce the potentially infinite variety of program mixes to a common dollar denominator, we (economists especially) should not lose sight of the particular quantitative and qualitative mix of inputs, which, after all, defines a program and which provides the information necessary to determine the ingredients of a program success or failure. On the other hand, program administrators should recognize that the notion "every program or particular project is different" can be pushed to the point of stifling all evaluations. Evaluations must be relative and comparative.

Most of the technical problems faced by the analysts on the input side are those of traditional cost accounting. We will confine our remarks to the two familiar and somewhat controversial problems of opportunity costs and transfer payments, which arise in nearly every manpower program. Both of these problems are most effectively dealt with if one starts by asking: What is the decision context for which these input measures are defined?

The most general decision context--and the one to which economists most naturally refer--is that of the productivity of alternative resource utilizations in society or the nation *as a whole*. In this case, one wishes to measure the cost of inputs in terms of the net reduction in value of alternative socially productive activities caused by the use of the inputs in this particular activity. Now, the value

of most inputs in terms of their alternative use will be more or less clearly indicated by their market price, but there are some inputs for which this will not be true. The most troublesome cases often concern the time of people. A well known example is the value of the time spent by students in school: since those over 14 or so could be in the job market, the social product (or national income) is less; therefore, an estimate is needed of what their earnings would be had they not been in school. (Such an estimate should reflect whatever amount of unemployment would be considered "normal.")

Sometimes the prices of inputs (market prices or prices fixed by the government) do not adequately reflect their marginal social productivity, and "corrected" or "shadow prices" are necessary. For example, the ostensible prices of leisure or of the housework of a wife are zero and obviously below their real price. By contrast a governmental fixed price of some surplus commodity is too high.

For manpower programs the best evaluation design would provide a control group to measure the opportunity costs of the time spent by the trainees in the program. Or, in measuring the value of the time of teenagers participating in a summer Upward Bound program, at least the question of market earnings foregone would be answered with a minimum of conjecture if control groups were available.

The definition and treatment of transfer payments also depend on the decision context of the analysis. From the national perspective money outlays from the budget of one program that are offset by reduced outlays elsewhere in society do not decrease the value of

the social product. When these outlays are in the form of cash payments or consumption goods, they are called transfer payments. An example is the provision of room and board for Job Corps trainees. Since it must be assumed that someone (their parents, themselves, or some welfare agency) would be meeting the costs of their room and board if they were not in the program, the provision of these services by the program reflects no *net* reduction in the value of alternative socially productive activities. Whoever was paying these costs before will be relieved of that burden and will spend the money thus saved on other goods and services. If there has been an actual *increase* in the value of food consumed by the trainee or in the quality of his housing, the net increase can be counted as a program input--a cost. But in general, it would be equal to the net increase in the value of food and housing consumed--a benefit.[16] To summarize, if these input costs are simply being *transferred* from one individual or agency to another individual or agency they either represent no real cost of resources of this program or they are a cost which is immediately offset by the benefit it yields to the recipient--remembering that the decision context is the general one which includes all members of society, with no one member receiving any different weight in the calculation of benefits.

In a narrower decision context, the accounting basis may shift; some input costs counted in the broader context are not counted in the narrower one and vice versa. One example of a narrow decision context--a favorite of people in government, but repugnant to most economists--is the vaguely defined "public budget." Alternatively

the decision context might be considered that of the "taxpayers' viewpoint" if the program participants and their families are excluded from the group considered as taxpayers. In this context the only costs that are to be counted are those that come from the public budget. Some of the examples we discussed above are now reversed. Presumably, most of the opportunity costs of a student's time spent in school is of no interest to the taxpayer since it is a "cost" which is not directly imposed upon the public budget. (A qualification is that the taxpayer should be interested in the taxes the student would pay if he were working.) By contrast the payments for the cost of room and board to a Job Corpsman, which was considered a transfer payment above, would now be considered an input cost from the "taxpayer's viewpoint." The fact that the trainee or his family is relieved of this burden would be of no interest since it would not be reflected in the public budget. However, if the costs of room and board had been met previously be a public welfare agency, then from the "taxpayer's viewpoint," the costs would not be charged to the Job Corps program.

It is not uncommon to see several decision contexts used in one analysis, and used inconsistently. For example, the post-training earnings improvement from participation in a Job Corps program are considered benefits. We all recognize, of course, that the earnings will be used mostly for consumption by the Job Corps graduate. But in the same study, his consumption during training (room, meals, and spending allowance), is not viewed as conferring benefits to the corpsman.[17] Or is it that the benefits should not count because

while in training, he is not considered a member of "our society?" We leave this puzzle to those who prefer these restricted decision contexts. There are other such examples and still other and more narrow decision contexts, such as that of a local government or of one project by itself. But it is probably clear that our preference is for the national or total societal perspective.

Program Outcomes. The problems of measurement on the outcome side of the evaluation problem are tougher to handle, and ex post evaluations of social action programs face particular problems because these outcomes are likely to involve behavioral relationships which are not well understood. It is particularly difficult to predict long run or permanent behavioral changes from the short run indicators revealed by the on-going or just completed program.

The outcomes we wish to measure from many social action programs occur months or years after the participants have completed the program. We can use proxy measures, which can themselves be measured during and soon after the program, but follow-up studies are clearly preferred and may in many cases be essential. A good deal depends on the confidence we have in the power of our theories to link the proxies or short-run effects (e.g., test scores, health treatments, employment experience in the short-run, etc.) with the longer run goals (longer run educational attainment, longevity, incomes, or all of these and perhaps other "softer" measures of "well-being"). It is a role for "basic research" in the social sciences to provide this type of theoretical-empirical information to evaluations, but we can also hope that the more thorough evaluation studies will contribute to our stock of "basic research" findings.

The problems of measuring longer run effects of a program and of conducting follow-up studies make up a long list, and most are familiar to administrators and analysts of social action programs. Some of these arose in our discussion of control groups where we noted the critical importance of identifying characteristics of respondents which would be related to the effects of the program and which may distinguish participants from the nonparticipants acting as a comparison group.

The problems of inadequate measures of variables and those of errors in the data are pervasive, particularly since the participants in the programs are often disadvantaged groups. Employment histories are checkered, making it difficult to determine the respondent's normal income, normal occupation, and other variables. Years of schooling completed may be a poor measure of educational attainment, police records may be an important source of employment difficulties, and so on. The above are but a few examples of the problems encountered in determining relevant data.

Measures of the status of a participant before entering the program usually come from the data gathered as part of the program intake procedure. A problem arises when potential enrollees are aware of criteria for program admittance for they may report inaccurate data in order to meet these criteria. Merely by sampling the data, the amount of inaccuracies can be approximately determined and appropriate correction factors can be devised.

The major obstacle to follow-up measures is the difficulty in locating people, particularly those from disadvantaged populations

who may be less responsive and who have irregular living patterns. The biases due to nonresponse may be severe, since those participants who are easiest to locate are likely to be the most "successful," both because of their apparent stability and because those who have "failed" may well be less responsive to requests to reveal their current status. One way around the costly problem of tracking down respondents for earnings data is to use Social Security records for participant and control groups. The rights of confidentiality may be preserved by aggregating the data.

Another problem in measuring outcomes, which also tends to be more talked about despairingly than coped with positively, is the category of external or third-party effects of the program. As a typical illustration consider a youth training program, which not only increases the earnings of the youths, but also reduces the incidence of crime among these groups, which benefits the community by way of less damage and through lower costs of prevention and rehabilitation programs. Another source of third-party effects are those accruing to the participant's family members, including those yet to be born. It is an open question, however, whether the problem for concern is the lack of measurement of these external effects, or the tendency by administrators and others (particularly friends of the programs) to exaggerate their likely importance and to count as external or secondary benefits those effects which, while benefiting some people do so at the expense of others.[18]

Concerning training and education programs, in particular, two types of effects that have received scant investigation are

structure or in various community institutions are assumed to be important because of the benefits or costs they ultimately provide for third-party individuals in the community. Thus, we are not proposing that the "community" be viewed as an "entity" separate from the individuals who comprise it. However, a separate focus on measures of community institutional changes appears necessary since the present state of our theories of community organization permit us little scope for anything except qualitative linkages between institutional changes and their effects on individuals in the community. We can, for example, consider better communication between the neighborhood populace and the police, school officials, or the employment service as "good things," either in their own right, as expressions of the democratic ethic, or because we believe that such changes will have tangible effects in safety, school achievement or better jobs.

Evaluations of social action programs may well have to deal with the problems of measuring variables that represent community effects even when such effects are not significant outcomes of a program. This need will arise when we have reason to believe that community institutions or aspects of the community structure are important independent or "control" variables that affect the program's objective. We have relatively well developed measures of some variables of the community structure, such as the components of a transportation system, but we are far less able to measure, for example, the degree of trust and rapport between the local branch of the State Employment Service and the poverty population in the community.

One major barrier to an adequate accounting of "community effects" is the scarcity of data pertaining to the community structure, although here we might argue, at the risk of revealing our prejudices or ignorance, that there is an overriding primary need for better theories of community structure and behavior. Without theory it is hard to know what facts or data we should be collecting.

The discussion of program outcomes again raises the problem of how to weigh and combine multiple objectives. Assuming that the separate objectives have been validly measured, the analyst might present the decision-makers with an array of multiple "effectiveness" measures and let them apply their own weights, explicitly or implicitly, to arrive at an over-all assessment, or he can use his own expertise and judgment to reduce the disparate outcomes to reasonably commensurable terms. The latter approach may be rationalized on the grounds that *some* such weighting scheme is inevitable and that an explicit method is better than a subjective one. For at least one aspect of commensurability--that of comparing goods and services that are identical except regarding *time*--the investment theory of economics provides a highly systematized method.

The Discount Rate. In general, society is not indifferent about whether a given outcome of a program is realized tomorrow or fifty years from now, and some attempt must be made to put outcomes and inputs on an equal time footing. The discount rate does this, and the controversy is over what the appropriate rate is. Without pretense that we are contributing anything original, we would simply like to report what we hope will be some clarifying views on the subject.

Since we argued earlier when discussing opportunity costs and transfer payments that our preferred perspective was that of the total society, rather than that of any single agency of the government or of the public fisc, we do not agree that the appropriate rate of discount is the cost to the agency or to the government of borrowing funds. This rate is unquestionably lower than that which stems from the societal productivity of alternative resource utilization.

It has often been argued that discount rates used for projects and programs in the public sector *should be* lower than those in the private sector. The basis for this argument is usually that people have a different rate of time preference for public than they do for private investments. If a dam or a health project in the public domain provides an effect 10 years from now rather than 5 years from now, we are less "unhappy", it is claimed, than we would be if a private investment in, say, a new apartment house pays off 10 years from now rather than 5 years from now. This argument is misleading because it confuses a difference in time preference with a difference in the value placed on the benefits. Whether the project is carried out in the public or private domain is surely not an important difference; it is rather a difference in the nature of the benefits. If we really believe that we make social judgments with a different (lower) time preference than private judgments, then we should use monetary and fiscal policy to force the rate of interest in the market down to the level of the social time preference and allow private and social projects to compete on an equal footing with respect to the rate of discount. The discount rate is, after all, simply a device which allows the time dimension of efficiency to be taken into account;

it reminds us that a project which can be completed in five years can yield a return during the next five years that we may reinvest and from which we realize further returns during the second five year period, whereas the 10 year project will only begin to pay off at the end of the second 5 year period. If we use different (and lower) rates of discount in the public sector, we attenuate this informational role of the discount rate and thereby give time-inefficient public projects an advantage over more efficient private activities.

A difference in the valuation of certain public as opposed to private activities is more accurately handled by giving a higher weight to the benefits of those public activities than they would carry if they were valued at the strictly equivalent market rates. If for example, a preventive health care program for a certain group of poor people raised their expected lifetime incomes by $100,000 we might well argue from the social point of view that this is worth more to us than the simple value of the increase in their private market productivity and multiply the benefit by, say, 1.2. This would leave unaffected the question of how to obtain such benefits most efficiently. Perhaps the basic objection to this procedure is the fear that if such weighting of benefits is made explicit, public administrators and decision-makers will not accept it, whereas manipulation of the discount rate gives them the desired result by a sufficiently obscure procedure which allows them to overlook the implicit weighting scheme. Using a lower discount rate is, in fact, strictly equivalent to multiplying benefits by some factor greater

than one, but there are practical reasons for eschewing the device of manipulating the discount rate. When a lower discount rate is used rather than a weighting of benefits, then there are arguments for using it for *all* public activities, and a subsidy is thereby provided to a whole host of activities which fall in the public domain merely by chance, tradition, or non-time-dependent efficiency considerations. Furthermore, differences in the social premium (or subsidy) above market valuations which we might agree to apply to the benefits from different types of programs call for many different discount rates, owing to the different durations and time patterns of the program. Either we will be juggling hundreds of discount rates, or we will fall back on a single rate for public projects which will fail to reflect the differences. Unfortunately, even if our comments up to this point are all correct, we are still unable to specify *the* correct discount rate, for this depends on what the appropriate governmental view of the risk element in its investment should be.[21] The market opportunity cost of capital is an obscure guide because of the multiple rates that exist in the face of varying risks. But even if a healthy allowance is made for the limited risk premiums involved in governmental investments, we should expect to be using rates of, at least, 7 or 8 percent. (Remember, that the perspective of the "total society" implies that before-tax rates of return on investments are the relevant measures of the opportunity cost of capital in private markets.) And if certain public ventures are especially worthy, we would again advocate that this should be reflected in the value of the benefits, not in any artificial suppression of the discount rate using the benefit-cost calculations.

In the face of discount rates which appear "high" by traditional standards in benefit-cost analysis of governmental programs, it may be worth pointing out that the *force* of these higher rates may be lessened in programs which involve investments in human capital, such as manpower training programs. If we take account of the "guaranteed" growth in per capita income in the economy--or, more exactly, in the increase in the "price" of "labor" (for a given quality level) the projected benefits to such programs as manpower training programs will increase--say at a rate of 2 percent. A short-cut allowance for this increase is to reduce the rate of discount used in the analysis by 2 percent and then project the constant levels of benefits which are available to us from the current data on wages and prices. This procedure has been used and defended elsewhere,[22] and here we should only like to point out that the basic source of this favorable treatment of human capital investments resides in part on the reasonable assumption of the relative flexibility of human beings to adapt to the diverse technological demands in an economy in which the quality and quantity of capital per worker is growing.

F. Organizational Problems

Timing and the Ability to Hold to Design. The effectiveness of evaluations of social action programs are highly dependent on the manner in which a number of organizational and administrative problems are handled. Although a thorough review of these problems is properly consigned to the literature of public administration, we feel it is important to discuss a few obstacles that can block even the best intentioned evaluator armed with the most sophisticated statistical and economic design.

In the beginning stages of planning an evaluation there are some important questions about the timing of the evaluation.[23] As social action programs are often innovative, it is not surprising that there is often a great clamor for an evaluation almost immediately after the program is begun. This is unrealistic since it takes some time for any program to settle down into "normal" operations, and program administrators are well aware of their tendency to progress along some kind of learning curve toward their maximum performance. In response to these points, it is sometimes argued that a "fair" evaluation of a program concept can only be undertaken a couple of years after a program has begun.

However, when the program to be evaluated is large scale and wide-spread, the organizational problems of setting up the evaluation can almost equal those of setting up a major project in the program. This means that the evaluative mechanism will need to be developed concurrently with the program organization. A failure to generate adequate information for analysis has been largely responsible for the paucity of meaningful evaluations of social action programs.

A related problem is that of insuring that programs hold to the initial design concept long enough to allow an evaluation to be completed. It is not uncommon to hear administrators complain that the evaluation they receive is well done but irrelevant, since the data used were taken from a period before certain fundamental changes were made in the program. The problem for the evaluator, then, is to complete his evaluation somewhere in the period between the "settling down" of the initial organization and the beginning of

fundamental shifts in the program process. (To some analysts this optimum period has begun to appear to be of about a week's duration). If program evaluation is to become an effective element in decision making it is important that there be an increased awareness both of the time it takes to set up and carry out an adequate evaluation and of the necessity of holding a program to a given design concept a sufficient length of time to allow such an evaluation process to be completed. And if we assume that the design of the evaluation provided for a wide range of variability in treatment variables, it is not likely to be irrelevant.

Internal Data Systems. The modernization of the management of public programs has led to an increasing interest in the internal data systems (sometimes called information systems) of programs. These systems are designed to facilitate the management of programs, including those functions we have characterized as "process evaluations" in Section II, but they can also be a great help for benefit-cost evaluations. There are several reasons, however, why an evaluator should not rely totally on an internal data system.

Administrators, especially at local levels, tend to place a low priority on data collection and analysis, and the result is that systems operators are seldom able to deliver on schedule the range of data which they originally promise. We have to recognize, also, that project operators sometimes have incentives to provide biased or simply manufactured data. Finally, internal data systems are notoriously inflexible, since the systems are usually designed with a limited set of users in mind. The result is that the analyst finds

it impossible to obtain disaggregations of these data or reaggregations by different sets of classifications. The importance of conserving micro-data has still not been generally appreciated.

For all of these reasons, the analyst is well-advised to supplement the internal data system with other information sources, perhaps by sampling from the system and perhaps through an outside source, such as the Social Security system. This procedure has the further advantage of liberating the internal data system from the burden of collecting for every participant all sorts of information vaguely believed necessary for "eventual" benefit-cost analyses with decisions about the selection of variables made by some one other than those who are planning the evaluation. For the purposes of the analyst, an internal data system which permits stratification and sampling may be all that is required.[24]

INTENTIONAL EXPERIMENTS: A SUGGESTED STRATEGY

Underlying the growing interest in evaluations of social action programs is the enlightened idea that the scientific method can be applied to program experience to establish and measure particular cause and effect relationships which are amenable to change through the agents of public policy. However, traditional methods in science, whether the laboratory experimentation of the physical scientists, the testing of pilot models by engineers, or field testing of drugs by medical scientists, are seldom models that can be directly copied, helpful though they are as standards of rigor.

In particular, evaluation designs patterned after the testing of pilot models, which correspond to "demonstration projects" in the field of social action programs, have been inadequate for both

theoretical and operational reasons. The present state of our theories of social behavior does not justify settling on a unique plan of action, and we cannot, almost by definition, learn much about alternative courses of action from a single pilot project. It is somewhat paradoxical that on the operational level the pilot model has failed to give us much information because the design has frequently been impossible to control and has spun off in different directions.

The combination of, first, loose administration of and rapid changes in the operation of individual projects and second, a large scale program with many heterogeneous projects (different administrations, different environments, different clientele, etc.), has led to the interesting view that this heterogeneity creates what are, in effect, "natural experiments" for an evaluation design. For economists, who are used to thinking of the measurement of consumers' responses to changes in the price of wheat or investors' responses to changes in the interest rate, the idea of "natural experiments" has a certain appeal. Certainly much of this paper has dealt with the problems and methods of coping with evaluations which attempt to take advantage of "natural experiments" within a program. But what should be clear from this discussion--and others before us have reached the same conclusion--is that a greatly improved evaluation could be obtained if social action programs were initiated in *intentional* experiments.

When one talks of "experiments" in the social sciences what inevitably comes to mind is a small scale, carefully controlled

study, such as those traditionally employed in psychology. Thus, when one suggests that social action programs be initiated in intentional experiments, people imagine a process which would involve a series of small test projects, a period of delay while those projects are completed and evaluated, and perhaps more retesting before any major program is mounted. This is very definitely *not* what we mean when we suggest social action programs as intentional experimentation. We would stress the word *action* to highlight the difference between what we suggest versus the traditional small scale experimentation.

Social action programs are undertaken because there is a clearly perceived social problem that requires some form of amelioration. In general, (with the exception perhaps of the area of medicinal drugs where a counter tradition has been carefully or painfully built up), we are not willing to postpone large scale attempts at amelioration of such problems until all the steps of a careful testing of hypotheses, development of pilot projects, etc. have been carried out. The practice, particularly in recent years, has been to proceed to action on a large scale with whichever seems--on reasonable, but essentially superficial, grounds--the best design at hand. We would suggest that large scale ameliorative social action and intentional experimentation are not incompatable; experimental designs can be built into a large scale social action program.

If a commitment is made to a more frankly experimental social action program by decision-makers and administrators, then many of the objectives we have advocated can be addressed directly at the

planning stage. If we begin a large national program with a frank awareness that we do not know which program concept is more likely to be most efficacious, then several program models could be selected for implementation in several areas, with enough variability in the key elements which make up the concepts to allow good measures of the differential responses to those elements. If social action programs are approached with an "intentionally experimental" point of view, then the analytical powers of our statistical models of evaluation can be greatly enhanced by attempts to insure that "confounding" effects are minimized--i.e., that program treatment variables are uncorrelated with participant characteristics and particular types of environments.

A less technical but equally important gain from this approach to social action programs is the understanding on the part of administrators, decision-makers, and legislators that if we are to learn anything from experience it is necessary to hold the design of the program (that is the designed project differentials in treatment variables) constant for a long enough period of time to allow for the "settling down" of the program and the collection and analysis of the data. *A commitment to hold to design for a long enough period so that we could learn from experience is a central element in the experimental approach to social action.*

The idea that social action programs should be experimental is simple, but we cannot be sanguine about the speed with which the full implications of this simple idea will be accepted by decision-makers and the public as a whole. The view that programs can be large scale *action* programs and still be designed as intentional

experiments has not been easy to get across, even to those trained in experimental methods in the social sciences, with its tradition of small scale research.

The emphasis on ex post evaluation is evidence of the fact that at some level legislators understand that social action programs are "testing" concepts. But it will require more explicit acceptance of the idea that some aspects of programs "tested" in action will fail before the full advantages of the intentionally experimental approach can be realized. It takes restraint to mount a program with a built-in experimental design and wait for it to mature before deciding on a single program concept, but we emhpasize that restraint does not mean small scale or limited action.

It is not unfair, we think, to characterize the approach to social action programs that has been taken in the past as one of serial experimentation through program failure. A program is built around a single concept, eventually it is realized that it does not work, so the program is scrapped (or allowed to fade away) and a new program and concept is tried. Certainly serial experimentation through failure is the hard way to learn. An intentionally experimental approach would allow us to learn faster by trying alternative concepts *simultaneously* and would make it more likely that we could determine not only *that* a particular concept failed, but also *why* it failed.

THE ACCEPTABILITY OF EVALUATION RESULTS

It does little violence to the facts to state that few decisions about social action programs have been made on the basis of the types of evaluations we have been discussing thus far in this paper. A

major reason for this, we feel, is an inadequate taste for rigor (or an overweening penchant for visceral judgments) by administrators and legislators and excessive taste for the purely scientific standards by academics. It often seems that the scholars conspire with the legislators to beat down any attempt to bring to bear more orderly evidence about the effectiveness of alternative programs; it is not at all difficult to find experts who will testify that virtually any evaluation study is not adequately "scientific" to provide a sound basis for making program decisions. There is a reasonable and appropriate fear on the part of academics that sophisticated techniques of analysis will be used as deceptive wrapping around an essentially political kernel to mislead administrators or the public. This fear, however, often leads to the setting of standards of "proof" which cannot, at present, given the state of the art of social sciences, or perhaps never, given the inherent nature of social action programs, be satisfied. The result generally is that the evaluation is discredited, the information it provides ignored, and the decision-maker and legislator can resume the exercise of their visceral talents.

A first step toward creating a more favorable atmosphere for evaluation studies is to recognize that they will not be final arbiters of the worth of a program. A positive but more modest role for evaluation research was recently stated by Kenneth Arrow in a discussion of the relative virtues of the traditional processes of public decision-making (characterized as an adversary process) and the recently developed procedure of the Programming, Planning, Budgeting System (characterized as a rationalistic or "synoptic process")[25]

Arrow advocated an approach in between forensics and synoptics.[26] He illustrated his argument by making an analogy with the court system, suggesting that what was happening through the introduction of the more rationalistic processes was the creation of a body of "rules of evidence." The use of systematic evaluation (along with the other elements of the PPBS) represents an attempt to raise the standards of what is admissible as evidence in a decision process that is inherently likely to remain adversary in nature. Higher standards of evaluation will lessen the role of "hearsay" testimony in the decision process, but they are not meant to provide a hard and fast decision rule in and of themselves. The public decision-making process is still a long way from the point at which the evidence from a hard evaluation is the primary or even the significant factor in the totality of factors which determine major decisions about programs. Therefore, the fear of many academics that poorly understood evaluations will exercise an inordinate influence on public decisions is, to say the least, extremely premature. But if standards for the acceptance of evaluation results are viewed in terms of the "rules of evidence" analogy, we can begin to move toward the judicious mix of rigor and pragmatism that is so badly needed in evaluation analysis.

The predominant view of the role of "serious," independent evaluations[27] (particularly in the eyes of harried administrators), seems to be that of a trial (to continue the analogy) aimed at finding a program guilty of failure. There is a sense in which this paranoid view of evaluation is correct. The statistical procedures used usually start with a null hypothesis of "no effect," and the

burden of the analysis is to provide evidence that is sufficiently strong to overturn the null hypothesis. As we have pointed out, however, problems of data, organization, and methods conspire to make clear-cut positive findings in evaluations difficult to demonstrate.

The atmosphere for evaluations would be much healthier if the underlying stance were shifted from this old world juridicial rule. Let the program be assumed innocent of failure until proven guilty through clear-cut negative findings. In more precise terms, we should try to avoid committing what are called in statistical theory Type II errors. Thus, an evaluation which does not permit rejecting the null hypothesis (of a zero effect of the program) at customary levels of statistical significance, may be consistent with a finding that a very large positive effect may be just as likely as a zero or negative effect.[28] "Rules of evidence" which emphasize the avoidance of Type II errors are equivalent to an attitude which we have characterized as "innocent until proven guilty." (We must frankly admit that, like court rules of evidence, this basic stance may provide incentives to the program administrators to provide data which are sufficient only for arriving at a "no conclusion" evaluative outcome.)

As a final conciliatory comment; when we talk about evaluation studies leading to verdicts of "success" or "failure," it should be recognized that we are greatly simplifying and abbreviating the typical results. Most social action programs are so complex in the variety of inputs and the multiplicity of objectives, that simple over-all judgments are not likely to lead to quick decisions to

dump programs. In combination with more detailed studies, the purpose of the evidence provided by the analysts will instead usually be to suggest modifications in the program--to shift the composition of inputs, perhaps to re-emphasize some objectives and de-emphasize others--and to suggest marginal additions or subtractions in the total scale of the program. It is worth emphasizing these modest objectives because the trust and cooperation of program administrators are indispensable to an evaluation of the program.

FOOTNOTES

1 As examples of the benefit-cost literature, see Robert Dorfman, ed., Measuring Benefits of Government Investments (Brookings Institution, Washington, D. C., 1965), and A. R. Prest and R. Turvey, "Cost-Benefit Analysis: A Survey," Economic Journal, December, 1965, v. 75, pp. 683-735. As examples of the evaluation research literature, see Edward A. Suchman, Evaluation Research (Russell Sage Foundation, New York, 1967), Donald T. Campell and Julian C. Stanley, Experimental and Quasi-Experimental Designs for Research (Chicago, Rand-McNally, 1966), G. H. Orcutt and A. G. Orcutt, "Incentive and Disincentive Experimentation for Income Maintenance Policy Purposes," American Economic Review, September, 1968, v. 58, pp. 754-72, and Harold Watts, "Graduated Work Incentives: Progress toward an Experiment in Negative Taxation," Discussion Papers Series, Institute for Research on Poverty, University of Wisconsin, 1968. For examples of the point of view of officials of governmental agencies, see William Gorham, "Notes of a Practicioner," and Elizabeth Drew, "HEW Grapples with PPBS," in The Public Interest, Summer, 1967, No. 8.

2 There does seem to be a developing literature in which the a priori benefit-cost estimates are compared with the ex post results for water projects. See Maynard Hufschmidt, "'Systematic Errors' in Cost Estimation in Public Investment," to appear in the Universities-National Bureau of Economic Research Conference volume, The Economics of Public Output. It may be that similar follow-up studies are being undertaken for defense projects--one can at least say that Congressional committees are determined to carry out their own follow-up evaluations on projects such as the TFX.

3 The Child Development Group in Mississippi (CDGM) was, for example, far more than just a Head Start program and the Tufts-Delta Health Center is more than just a health program. Jobs were generated and incomes were raised which brought about changes in the power structure of the community. For two revealing journalistic accounts in Life magazine, see: David Nevin, "Struggle That Changed Glen Allan," Vol. 63, Sept. 29, 1967, pp. 108-112, and Richard Hall, "A Stir of Hope in Mound Bayou," Vol. 66, March 28, 1969, pp. 66-81.

4 The characteristics of Community Action Programs and the problems they create for operating and evaluating the programs are forcefully discussed by Daniel P. Moynihan, in his book, *Maximum Feasible Misunderstanding: Community Action in the War on Poverty*, New York: Free Press, 1969.

5 We are indebted to Thomas K. Glennen, RAND Corporation, for his ideas on this point.

6 Briefly, the OIC concept combines elements of training, job development (often aided by pressure tactics against employers), and a psychological up-lifting of the participants which is conducted with an ideology of militancy and participatory democracy.

7 The Work Experience program consisted of public employment of welfare recipients and other adult poor under Title V of the Economic Opportunity Act. Only minimal training was offered, but it was hoped that work-for-pay would, by itself, provide a springboard to self-sustaining employment in the private market.

8 U. S. Congress, House Committee on Ways and Means, Community Work and Training Program. 90th Congress, 1st Sess., House Document No. 96 (Washington, D. C.: U. S. Government Printing Office, 1967).

9 Worth Bateman, "Assessing Program Effectiveness," Welfare in Review, Vol. 6, No. 1, January-February 1968.

10 One range for a confidence interval of special interest is almost always that which includes zero for its lower limit (thinking now of a social action program that has some positive effect), so that the investigator is able to test the null hypothesis that the program makes "no difference." This is conventional, and so is the practice of measuring the quantitative magnitude of the effect when the null hypothesis is rejected. We should not overlook, however, the information about the range of quantitative effects of variables even when the their confidence intervals include zero and when, therefore, the null hypothesis of "no effect" is accepted. Clearly, we would want to know that the interval was, say, -$5 to $455 rather than -$455 and $5. Furthermore, there are any number of situations when we should be interested in weighing the seriousness of negative effects with the benefits from, possibly, very large positive effects. Put in other terms, zero is bracketed by -$5 to +$5 as well as by -$500 to +$500, and there may be situations in which it is important to distinguish between the two cases.

11 In the absence of an ex post evaluation, such a priori analysis would be useful in assessing the general feasibility of the project.

12 We may well have in mind attempting a number of different programs that are radically innovative and for which our a priori notions predict either spectacular success or complete failure. A

program to cure narcotics addiction might be such a program. Given the costliness of properly designed evaluation schemes, we might justify pushing ahead with the programs without waiting on formal evaluation procedures in the hope that even "casual observation" will render a valid verdict of the program.

13 An important point to be remembered is that, for any given amount of resources available for an evaluation study, there is a trade-off between an allocation of these resources for increased sample size and allocation for improved quality of measurement, which might take the form of an expanded set of variables, improved measures of variables, or reduced attrition from the sample. Too often we have witnessed a single-minded attachment to larger sample sizes, probably stemming from the analyst's fear that he will end up with "too few observations in the cells" of some only vaguely imagined cross-tabulation. This fear should be balanced by an awareness both of the rapidity with which marginal gains in precision of estimates decline with increases in "medium size" samples and of the extent to which a theoretically justified multiple regression model can overcome some of the limitations which cross-tabulation analysis impose on a given-sized sample.

14 See the vigorous defense of the experimental method in social action programs in: Guy H. Orcutt and Alice G. Orcutt, op. cit.

15 This assumption will strike some readers as too positivistic, too restrictive to "things measurable," and too oblivious to the unmeasurable and subjective variables. Let us say in defense of this assumption only that it is a "working assumption" that permits us to discuss an important region of evaluation which covers the measurable portion, that it is desirable to expand this region and, therefore, to narrow the area left for subjective judgments, and that, in any case, the objective portion is necessary to an improved over-all judgment that spans both measurable and unmeasurable inputs and outputs of a program.

16 When the program produces an increase in consumption of goods and services, the treatment of these transfer payments can become more complicated if we do not assume that the goods and service have a value to the recipients equal to their cost. See A. A. Alchian and W. R. Allen, University Economics (Wadsworth: Belmont, California, 1967, Second Edition) pp. 135-140 for an extended discussion.

17 For just one of many examples of this type of treatment of transfer payments see, "The Feasibility of Benefit-Cost Analysis in the War on Poverty: A Test Application to Manpower Programs," prepared for the General Accounting Office, Resource Management Corporation, UR-054, December 13, 1968.

18 For a notable exception to the absence of attempted measurement of the type of third-party effects discussed above, see Thomas I. Ribich, Education and Poverty (Washington, D. C.: The Brookings Institution, 1968). Ribich's study also gives us some evidence of

likelihood of relatively small quantitative magnitudes of these effects. A rather free wheeling listing of third-party effects runs the risk of double counting benefits. For example, although other family members benefit from the better education earnings of the head of the household, we should not forget that had the investment expenditure been made elsewhere, even if in the form of an across-the-board tax cut, other family heads would have had larger incomes, at least, with resulting benefits to their families. In his examination of cost-benefit analysis of water resource developments, Roland N. McKean gives an extended discussion of the pitfalls of double counting. See his Efficiency in Government Through Systems Analysis (New York: John Wiley and Sons, Inc., 1958), especially Chapter 9.

19 An exceptionally good discussion of negative external effects, including disruption to the community structure, is contained in Anthony Downs, "Uncompensated Non-Construction Costs Which Urban Highways and Urban Renewal Impose on Residential Households" which will appear in a Universities-National Bureau of Economic Research Conference volume entitled, Economics of Public Output. The literature on urban renewal and public housing is extensive and too well known to require listing here.

20 For an excellent discussion of many of these issues see Joel F. Handler, "Controlling Official Behavior in Welfare Administration," The Law of the Poor, ed., J. tenBroek (Chandler Publishing Co., 1966). (Also published in The California Law Review, Vol. 54, 1966, p. 479.)

21 Compare Kenneth J. Arrow, "Discounting and Public Investment Criteria, in Water Research, ed., A.V. Kneese and S. C. Smith, (Johns Hopkins Press, Baltimore, 1966), pp. 13-22, and Hirshleifer, De Haven, and Milliman, op.cit. on this point.

22 Glen G. Cain, "Benefit/Cost Estimates for Job Corps." Discussion Papers, Institute for Research on Poverty, The University of Wisconsin, Madison, Wisconsin, especially pp. 12, 17-18, 39-42. See also, Gary Becker, Human Capital, National Bureau of Economic Research, Number 80 (New York: Columbia University Press, 1964), especially p. 73.

23 We are indebted to discussions and correspondence with Thomas K. Glennan, RAND Corporation, for many of the ideas in this section.

24 It has often proved surprisingly difficult to convince program managers that for the purposes of evaluation small samples of data are perfectly adequate and that, in some cases, data gathered on the entire "universe" of the program are cumbersome or costly to manipulate, are notoriously error-laden, and generally add little additional useful information.

25 For a more complete discussion of this terminology, see Henry Bowen, "Recent Developments in the Measurement of Public Outputs," to be published in a Universities-National Bureau of Economic Research Conference volume, The Economics of Public Output.

26 Remarks by Kenneth Arrow during the NBER conference cited in the previous footnote.

27 We mean here to exclude the quick and casual sort of evaluations, mainly "in-house" evaluations, that more often than not are meant to provide a gloss of technical justification for a program.

28 Harold Watts has stressed this point in conversations with the authors. See Glen G. Cain and Harold W. Watts, "The Controversy about the Coleman Report: Comment,"Journal of Human Resources, Vol. III, no. 3, Summer, 1968, pp. 389-92, also, Harold W. Watts and David L. Horner, "The Educational Benefits of Head Start: A Quantitative Analysis," Discussion Paper Series, The Institute for Research on Poverty, University of Wisconsin, Madison, Wisconsin.

END