THE SOCIAL EXPERIMENTS: STATISTICAL ISSUES AND THE ROLE OF STATISTICIANS

Charles E. Metcalf, Mathematica Policy Research [1]

A. INTRODUCTION

In the past ten years the methodology of controlled experimentation has taken firm hold as a focal point of analyses of major changes in social programs. Beginning with the New Jersey negative income tax experiment, there have been several major controlled experiments--with randomized assignment of participants to treatment programs and a control group--in the areas of income maintenance, housing allowances, health insurance, and employment programs. Furthermore, the existence of such experiments has had a major impact on the methodologies used for program evaluations not employing randomly selected control groups.

A major apparent concern to social scientists outside the economics profession is not so much the experiments themselves, but their own lack of involvement in the design and execution of the experiments. Critiques of the New Jersey experiment by sociologists have stressed how the analysis would have been improved by the inclusion of more (or higher quality) sociologists. The American Statistical Association has now scheduled a session on "the role of the statistician in social experimentation" and further compounded the issue by inviting an economist to deliver a paper on the topic.

While a broad range of the social sciences was represented in the major social experiments--sociologists, psychologists, political scientists, and statisticians--it is true that economists have played a dominant role in the recent growth of social experimentation. With regard to issues of sampling and statistical methodology, some statisticians were utilized, but there was a clear tendency for economists to call upon econometricians with training in statistics rather than upon individuals regarding themselves as statisticians.

These developments in the social experiments were an extension of tendencies already apparent in economics as a profession. Economists have been more heavily involved in quantitative measurment and hypothesis testing than have sociologists, and in the process econometrics became a sophisticated "in house" form of applied statistics. The earlier income maintenance experiments in particular were primarily intended to test hypotheses about economic behavior already well developed in economic theory--but with a data set not subject to the limitations inherent in the historical data commonly used. Despite the methodological departure into the realm of controlled experiments, the social experiments have retained both the perspective and the analytic approach of economists.

The sociologists' lament concerned economists' perspective about the objectives of the experiments--i.e., whether the right questions were being asked. The absence of participation by statisticians, if in fact true, relates to whether executers of the experiments availed themselves of available expertise when dealing with the statistical problems inherent in social experimentation.

There are four major areas where scientific expertise generally belonging in statistics as a disciplinary classification was required for execution of the social experiments: (1) establishment of the sampling frame, (2) experimental design, (3) empirical estimation and hypothesis testing, and (4) extrapolation of empirical results to quantitative measures relevant for policy decisions. While these four areas are closely interrelated I would like to discuss each area in turn, stressing the statistical issues involved, and speculating about what the impact of greater involvement of statisticans might have been. I shall begin by specifying a prototype design model with a simplified structure--having clear statistical implications but corresponding to none of the experiments actually performed. With specific examples from the New Jersey Experiment and from the other major experiments, each deviation from the prototype model can then be judged both in terms of the advantages claimed to justify it and the statistical problems associated with it.

The remainder of this introduction summarizes the key features of some of the major social experiments to date and outlines the prototype design model. Sections B through E address the four areas of scientific expertise cited above, while section F provides some concluding remarks.

The Major Social Experiments

The following research projects all utilized some form of randomized assignment of individuals or households into a treatment group, who were eligible for participation in the social program being evaluated, and a control group who were not eligible for program participation but whose behavior was observed. The list is not meant to be exhaustive; rather it includes the major income maintenance experiments plus illustrative social experiments in other areas with which I have some familiarity.

The New Jersey Income-Maintenance Experiment, funded by the Office of Economic Opportunity beginning in 1967, was the first and most widely scrutinized of the social experiments.[2] Several forces joined to bring about the experiment--increasing advocacy of negative income taxation as a viable policy option, combined with continuing concerns about the potential effect of universal income maintenance on work incentives; the feeling in both the scientific and political communities that evidence on work incentive effects obtained to date with non-experimental

techniques was insufficient; rising interest in social experimentation as a viable research option; and a proposal to undertake such an experiment by Heather Ross, then a Ph.D. candidate in economics at M.I.T.[3] The primary research question to be addressed was the effect of a negative income tax (N.I.T.) on the labor supply of families with a prime age, able-bodied male. A sample of 1,357 such families was drawn from five urban centers in New Jersey and Pennsylvania,[4] and 725 were allocated to a variety of NIT plans for a period of three years. Including design, staggered payment periods, and subsequent analysis, the experiment lasted seven years and cost $7.6 million dollars.

The Rural Income Maintenance Experiment applied a similar experimental design to 800 families in rural areas of Iowa and North Carolina. Its smaller sample size was fragmented by the designation of 100 sample points as households with a female head, and 100 as households with a head over 55 years of age; and by the stratification by farm/non-farm status as well as by geographic location.

The Gary Income Maintenance Project utilized a predominantly black sample of 1,600 families in Gary, Indiana, over half of which were families with female heads already receiving Aid for Dependent Children.

The Seattle/Denver Income Maintenance Experiment is the last of the U.S. income maintenance experiments and the most ambitious. It has a larger sample (5200); varies the duration of payments (3, 5, and 20 years) to test long-run program effects; includes job counseling and training for a fraction of the sample; and utilizes a non-linear tax rate structure for some treatment plans.

A flurry of interest in experimentation in Canada almost led to a similar array of Canadian experiments, but only the Manitoba Minimum Annual Income Project got off the ground, with a New Jersey style experimental design in Winnipeg and a "saturation" site in Dauphin, Manitoba.

Outside the realm of income maintenance, a series of three Housing Allowance experiments made payments to low income households either conditional on meeting minimum housing standards, or determined as a percentage subsidy of rental expenditures. The Health Insurance Study involved randomized assignment of individuals to different health insurance plans.

Supported Work provides transitional employment to ex-addicts, ex-offenders, AFDC mothers, and youth groups believed to have special problems adapting to conventional jobs. Objectives of the program include measurement of post-program employment, criminal behavior, drug use, and within-program productivity. A simple random assignment to program and control group status is made among applicants meeting eligibility standards. Colorado Monthly Reporting examines the effects of converting a random sample of Denver AFDC recipients to a computerized payment system based on the previous month's income, rather than the conventional method of a needs determination every six months; in addition, the entire caseload of Boulder County was placed on a monthly reporting system.

The Prototype Model

To facilitate the evaluation of the four critical areas where decisions relating to statistical methodology were made in the social experiments, the following prototype model can be a useful point of departure. The prototype model involves four explicit steps:

(1) The drawing of a simple random sample of all households in the U.S.;

(2) Simple random assignment of a program treatment to half the sample, with the remaining sample designated as the control group;

(3) Direct comparison of treatment and control groups to measure program effects, with simple difference-of-means statistical tests; and

(4) The interpretation of results as measuring directly what would happen with full scale adoption of the program.

While the above model is appealing in its simplicity, there were felt to be persuasive reasons for making major departures from the prototype at all four steps.

B. THE SAMPLING FRAME

None of the major experiments utilized a simple random sample of the sort envisioned in the prototype model. Major issues debated by designers of the experiments included (1) restriction of the sampling universe to policy-relevant subsets of the population; (2) the adoption of dispersed sampling versus implementation of "saturation" experiments; (3) the use of a random national sample vs. "test bores" of the population in a small number of sites; and (4) sampling procedures within sites.

The first major departure from the prototype model involved the truncation of the sample universe to include only families meeting certain income-eligibility and demographic criteria. Most of the programs being considered were targeted to particular segments of the population. For example, the NIT is targeted to low income households, even though all members of the population are eligible if their incomes fall into the relevant range. Thus the decision was made to sample only low income households rather than to observe large numbers of higher income people who were unlikely to receive program benefits. This truncation of the sample by income level was the most severe in the case of the New Jersey experiment, which included only families with incomes below 150% of a commonly used poverty income definition. The sampling universe for the New Jersey experiment was further restricted to include only families with a prime-age, able-bodied male. Since the labor force response of

such families was believed to be pivotal in the evaluation of a universal income maintenance scheme, it was decided to concentrate efforts on testing hypotheses about one group rather than to spread resources across heterogeneous family types.

Such sample truncation led to at least two problems. First of all, since income is under the partial control of household members through labor market decisions, the truncation variable was closely related to behavioral responses being measured in the experiment. It has been well established that when a sample truncated to restrict the domain of a variable is used to estimate determinants of that variable, conventional regression techniques will lead to biased results.[5] Secondly, it was realized ex-post that the truncation process eliminated the possibility of measuring program effects of major interest. Specifically, there were many two earner families who would receive NIT payments if one of the earners quit his or her job, but who had incomes in excess of 150% of the poverty level so long as both jobs were retained. Thus the severe truncation of the New Jersey experiment prevented a proper test of this response by excluding such families from the sample; available evidence suggests that this would be one of the largest sources of work reduction as a result of the NIT.

Subsequent experiments alleviated the problem by truncating the sample at a much higher level of income, and by including a broader range of demographic groups. At least one experiment went too far in the direction of sample heterogeniety. With a sample of only 800, the Rural experiment included female and aged household heads as well as families with prime-age males and stratified by geographic region and farm/non-farm status. The result was a sample with very little power for testing behavioral hypotheses.

The prototype model implies the choice of a dispersed sample rather than a saturation sample. One might expect that an individual's response to a program in which he participates as part of a random sample may differ from one where all individuals like him in the same community are subject to the same program. Also, saturation may be required to measure community effects on non-program participants; to observe major economic responses to the program (e.g., the effect on housing supply of a major housing subsidy program); and to evaluate the operational feasibility of implementing a full scale program. The idea of a saturation experiment including all program-eligibles in a random sample of localities has been frequently discussed but never implemented. The rejection has usually been on cost grounds. Several of the social experiments did, however, include saturation of selected sites without the explicit use of corresponding control groups--in particular, the Housing Allowance Supply Experiment and portions of the Manitoba and Colorado Monthly Reporting experiments.

The rejection of a random national sample in favor of a "test bore" sample in a geographically limited area was one of the more controversial decisions, made first in New Jersey and replicated in subsequent social experiments. Against the obvious loss of all statistical power for national extrapolations were placed the following advantages of a "test bore" sample: the (ultimately dominating) issues of cost and administrative feasibility, and the ability to test hypotheses against a homogeneous background environment. Given the limited resources available for experimentation, increasing the power of within-sample hypothesis testing was felt to be more important than representativeness of the sample. Most people involved with the experiments continue to believe that this decision was a correct one, at least given the information available at the time. I suspect that if statisticians rather than econometricians and policy analysts had led the movement to social experimentation, however, the case for a national sample would have been more forcefully defended.

The lack of a statistician's orientation also had an impact on the sampling procedures within sites, particularly in the New Jersey experiment. Because the yield of low-income households from New Jersey screening interviews was much lower than anticipated, cost considerations required that enumeration of the sampling frame be limited to census tracts with a high incidence of low income households. Thus poor families in low density areas had a zero probability of selection, and this fact resulted in the major unanticipated feature of the New Jersey sampling frame: the predominance of blacks and Puerto Ricans, and the resulting foray to Scranton, Pennsylvania in search of poor whites.

C. EXPERIMENTAL DESIGN[6]

In our prototype experiment, we observed the effects of a single policy change and interpreted the results directly. Some of the social experiments--Supported Work, for example--adopted designs similar to the prototype. The New Jersey experiment and most of its successors deviated from the prototype, however, both by including a multiplicity of experimental treatments and by adopting a complicated method of assigning sample households to specific treatments.

The case for a more complex experimental design rests on three major arguments:

(a) Policy interest is focused not on a single, known program but rather on a range of programs with similar characteristics.

(b) The experimental environment cannot provide a direct test of the relevant policy issues; thus the experimental design must provide the necessary information for extrapolating the results to the appropriate environment.

(c) An efficient experimental design should reflect prior knowledge about the structure of hypotheses to be tested and about differential costs of alternative experimental treatments.

From the nature of the reasons given, it should be fairly apparent that decisions regarding experimental design, hypothesis testing, and extrapolation of results are closely intertwined. Despite this, I shall maintain the fiction of distinguishing among the three areas, and outline the principles of experimental design developed for the New Jersey experiment.

Let us first consider a "design space" of potential program treatments as a range of testable programs of direct policy interest. If, for instance, we knew that our policy choice were limited to a single negative income tax plan versus no plan at all, we might limit our design space to a single plan plus a control group, as in the prototype design. If our objective were to choose among three plans, we might opt for a design space with three corresponding plans in addition to a control group. Such a design would permit a comparison of behavioral responses between any two treatments, and between a single treatment and the control group.

An increase in the number of treatments given a fixed budget or total sample size obviously reduces the number of observations per cell, and thus the precision of any pairwise test. It is desirable to develop some method of exploiting similarities among responses to alternative treatments not only to alleviate the loss of precision involved in testing multiple treatments, but also to make statements about behavioral responses to similar treatment plans not explicitly included in the experiment. This latter issue can be of major importance if the set of policies having potential policy interest shifts during the course of the experiment.

The notion of similarities among treatments suggests an alternative view of the design space as a range of program characteristics that affect household behavior, with a range of plan characteristics rather than merely specific treatments being of direct policy interest. In the case of the New Jersey experiment, each NIT plan was defined by a specific combination of income guarantee, G, and tax rate, t. The motivation behind restating each treatment in terms of characteristics influencing behavior came from the added assumption that behavioral responses vary in some continuous manner with variation in plan characteristics. If the relationship of behavior to variations in G and t can be approximated by a continuous response function of known (maximum) dimension, a design space including values for G and t at the extremes of the range of potential policy interest, plus sufficient interior values to identify the assumed response function, provides information about a complete continuum of policy options rather than simply a limited set of specifically tested alternatives. Correspondingly, precision in the estimated response at a specific G and t combination is derived not only from observations at that point, but from all observations relevant for identifying the "response surface." Because extrapolation of the effects of G and t combinations beyond the extreme observed variations can be done (if at all) with less confidence, the emphasis in this framework shifts to specifying the extremes of potential policy interest rather than the points of greatest direct policy interest. That is, even if we are most interested in obtaining information about central regions of our "policy space," this interest may be best served in an experiment stressing treatments at the fringes of our range of interest.[7/]

Once we think of obtaining information about a policy space in terms of testing hypotheses relating to household responses to program characteristics, plans to be included in the design space and those of direct policy interest may no longer coincide. Even if we know with certainty the policy alternatives to be considered, the optimal experimental design could, under some circumstances, not only exclude certain treatments of direct policy interest, but also include other treatments not among the set of policy alternatives.

The range of treatment plans can also be defined in terms of characteristics required to extrapolate results to a nonexperimental setting (see Section E). For example, the income maintenance experiments were of limited duration, whereas the programs of policy interest are presumably permanent. In order to permit the appropriate projections to be made, the Seattle/Denver experiment systematically varied the duration of its treatment programs from a minimum of three years to a maximum of 20 years.

The above discussion indicates that a careful consideration of experimental objectives could result in a design space that differs from a simplistic statement of programs of direct policy interest. Similarly, a sample allocation process that takes into explicit account both specific experimental objectives and budgetary and other constraints may lead to a violation of some commonly proposed principles relating to orthogonality of the sample design--these principles involve relationships (1) between plan assignments and household attributes and (2) among plan characteristics or variables to be included in an estimated behavioral model.

Given a situation where the design space is defined as a set of policy alternatives and where a decision has been made regarding the number of households to be assigned to each plan, it is often proposed that households for each cell be chosen by a self-weighting random sampling procedure. Even if the aggregate sample is to be stratified by certain household attributes, the view holds that the stratification characteristics should not influence the probability of assignment to a specific plan. That is, orthogonality of plan and stratification characteristics would be maintained so that simple comparisons could be made across plans.

Orthogonality is also typically stressed as a desirable feature of sample allocations across design spaces dimensioned in terms of plan characteristics because it permit hypotheses concerning a single characteristic to be tested without having to control for variations in

other plan and stratification characteristics. Indeed, orthogonality is an optimality condition for a class of problems often discussed in the design literature.[8]

Consider a case where the objective of the experiment has been defined as measuring experimental response relative to the control group for each of several treatments. In this case a regression form of an analysis of variance framework suggests itself where household behavior is viewed as a linear function of a set of dummy variables (one for each plan), and where the goal is to obtain accurate estimates of the coefficients associated with the differential effect of the experiment at each design point. If we specify the objective to be the minimization of a weighted sum of coefficient variances given a budgetary restriction, the optimal allocation of households could correspond to the uniform distribution across plans proposed above--*if* equal weight is attached to each variance term *and* costs per observation are identical across plans.

This latter condition is violated in the case at hand, since an intrinsic feature of NIT plans is that costs per observation vary systematically with plan characteristics--namely, the guarantee level and the tax rate. Starting from an initial uniform allocation where sample points of differing costs make the *same* marginal contribution to the experimental objective, the efficiency of the design could be improved by surrendering some expensive observations for a larger number of cheaper ones.

This latter result strongly influences the allocation of households to the control group; which is far less expensive per observation than the experimental cells. For instance, in order to measure with minimum variance the differential behavior between a control group and a single experimental cell in a situation where the cost per observation for experimentals is nine times that of controls, 75 percent of the sample should be assigned to the control groups and only 25 percent to the experimental treatment. Compared with an allocation of 300 treatment observations and 900 control observations, moving to equal cell sizes (360 each) would increase the variance of our estimate by 25 percent. Given the cost assumption which generated the three-to-one ratio between controls and experimentals, 75 percent of the budget is still expended on experimental plans. Thus, other things equal, a more expensive plan would be allocated a smaller number of observations in the optimal design, but would command a larger share of the experimental budget.

Cost differentials also play a role in the decision whether or not to stratify the sample by household characteristics. (The other major consideation is whether identification of differential responses by household characteristics plays an explicit role in the experimental objective.) If the population of interest were stratified by characteristics which affected costs per observation (e.g., family size or income), and if the experimental objective were to estimate the mean population response to a given treatment, the optimal strategy would be to oversample in those subgroups for which information could be obtained more cheaply.

It should be apparent that accounting for cost differentials in the sample allocation process is sufficient to destroy orthogonality between experimental variables and population characteristics as an optimality condition. Because the cost differential between experimental and control observations depends on household income, for example, the probability of assignment to a particular cell would no longer be independent of income. Basic principles of randomization are retained, however, if all households within a single stratum (that is, households identical in terms of stratification characteristics) face the same set of assignment probabilities.

The sample allocation models used in the income maintenance experiments went further than simply to account for variations in observation costs. The Conlisk-Watts model[9] which formed the basis of sample allocations in the income maintenance experiments has four major components:

(1) an assumed structural relationship, specified as a regression model, relating behavioral responses to treatment and household characteristics;

(2) a "design space" relating each treatment plan and household stratification to the structural model;

(3) an objective function, providing the measure by which the desirability of a design allocation is judged; and

(4) a total budget constraint and a vector specifying the cost per observation at each point.

Given the above information, the design problem is then to choose that distribution of households across design points which optimizes the objective function. Like the cost constraint, the objective function is capable of introducing factors which imply that nonorthogonality is a desirable feature. While its specific form may vary, in general we wish to minimize a weighted sum of variances associated with a vector of linear combinations of regression coefficients. The solution to the design problem specifies the number of households from each stratum to be allocated to alternative treatments; individual households are then randomly assigned according to the selection probabilities implicit in the solution.

If the Conlisk-Watts model begins with a correctly specified structural relationship, it can be a valuable tool in increasing the efficiency of an experimental design. It has been criticized, however, by those not wishing to let prior structural assumptions (which may be incorrect) condition the experimental design, and by the complexities it introduces in the use of experimental data for hypothesis testing. Some

of these issues will become apparent in the next section.

D. EMPIRICAL ESTIMATION AND HYPOTHESIS TESTING

The prototype model focused on the testing of a simple direct hypothesis concerning experimental effects. The experimental designs used for the income maintenance experiments were intended to accommodate more complicated hypothesis tests involving both variations in program characteristics and extrapolations to nonexperimental settings. The sample allocation was intended not only to permit the testing of these more complex hypotheses, but also to promote the precision of the intended tests.

While the samples for the NIT experiments were drawn according to the fundamental randomization principles necessary for applying conventional techniques of statistical inference, the design process was permitted to determine the choice and frequency of applied treatments and to choose probabilities of selection in alternative purposes and must be accounted for in designing methods of analysis.

The first point to be made is that the experimental design places limitations on hypotheses which can be tested. The New Jersey experiment was designed to vary controlled characteristics in a finite number of dimensions, and in such a way as to permit efficient testing of hypotheses related to a particular regression model. Alternative hypotheses may be tested, so long as the design space has sufficient dimensions to accommodate the tests. Such tests will be less efficient than if the experimental design had been established with those tests in mind. Thus the prototype design required hypothesis tests to be simple, but permitted more powerful tests of simple hypotheses than the designs used in the income maintenance experiments.

Secondly, the sample allocation process in the income maintenance experiments created a correlation between some household characteristics and form of program treatment. Thus simple bivariate relationships and comparisons of group means no longer have the direct interpretive value they would have had with orthogonal designs. For example, a simple test comparing mean earnings of families in a particular plan with those in the control group may be contaminated by the fact that family income influences the probability of assignment among treatments. That is, households in a particular plan and in the control group may be systematically different in terms of stratification characteristics, and simple group comparisons do not permit one to distinguish between the effects of plan and stratification variables on observed behavior.

This problem can be rectified by explicitly incorporating all stratification characteristics into the hypothesis test--either by controlling for all stratification characteristics in performing the test, or by making (and presumably defending) the assertion that the response being observed is independent of the stratification variables in question. Generally speaking, stratification of a sample by variables appearon the right-hand side of a regression model has no effect on the interpretability of coefficients or test statistics associated with that model.

The use of complicated experimental designs introduces definite risks that the prototype model avoided. It is necessary to specify a structural relationship between the behavioral response of interest and all variables used for stratification purposes in the design process. If this structural relationship is subject to specification error, the resulting experimental inferences may be incorrect. The prototype model, on the other hand, was more conducive to tests of experimental effect without knowledge of the underlying structure.

A third issue relates to projecting population estimates from results based on the experimental sample. The premise of conventional sampling theory--that a self-weighted random sample constitutes an unbiased representation of the population of interest--is not applicable to a situation where we induce behavioral responses in an experimental setting and requires an explicit theory for extrapolating to a nonexperimental situation.

Once we have confronted this situation, we may wish to translate measures of behavior for the experimental sample into unbiased estimates of what these measures would have been for a self-weighted sample of the population. The proper procedure involves a simple reweighting of the sample measures. There is a fundamental rule to be followed in this process, however, which is frequently violated: first estimate behavioral relationships on the raw sample, then reweight the distribution of point estimates where appropriate.

The reverse procedure of weighting observations prior to testing hypotheses, while equivalent for the direct calculation of variable means, results in incorrect estimation procedures in a regression framework. Consider an example in which labor supply is correctly specified as a linear function of the guarantee, the tax rate, and normal earnings, with a homoschedastic error term. Given these assumptions the appropriate estimation procedure, independently of how the distribution of observations by normal earnings corresponds to that of the population of interest, is unweighted least squares. To weight the observations would introduce heteroschedasticity in the error term and lead to an inefficient estimation procedure. If the error term in the raw regression is heteroschedastic, the weighting of observations and regressors (including the intercept) is an appropriate correction procedure, but these weights would bear no relationship to those involved in constructing population estimates.

Similar care must be taken in using experimental data sets for estimating behavioral relationships unrelated to the experiment. In particular, attempts to estimate behavioral

relationships involving stratification variables as dependent measures must utilize special estimation techniques.[10]

Finally, some general problems of statistical methodology related to hypothesis testing should be mentioned. In addition to the standard analytic problems associated with panel survey data--e.g., the need to deal with auto-correlated stochastic terms and with non-response bias and sample attrition--the fact that behavior has been experimentally manipulated creates special problems. Most of the experiments have been confronted with differential sample attrition rates by program status. Further difficulties are created when the structural models being tested require proxy variables such as "normal income," frequently essential in models of household economic behavior. On the one hand, it is hard to obtain a proxy free of induced experimental effects from the data for treatment households; alternatively, the construction of proxy variables from the same control group data used for making treatment-control comparisons can lead to small sample bias in constructing certain types of hypothesis tests.[11]

In summary, the designers of the income maintenance experiments deviated substantially from simple models in an effort to make the same design responsive to the structure of the hypotheses being tested. The cost imposed by this procedure was immense in terms of complexity imposed on the hypothesis tests and in terms of subtle analytic pitfalls created in the process. The net value of these design efforts is a subject of continuing dispute, with economists often taking a different position from observers in other disciplines.

E. EXTRAPOLATION TO NON-EXPERIMENTAL SETTINGS

The problem of extrapolation to non-experimental settings lies at the center of what makes the design of experimental samples different from the traditional practices of survey sampling. In survey samples we wish to obtain information about existing population characteristics without contaminating either behavior or household responses by the choice of survey methods. So long as such contamination can be avoided, well established random sampling procedures permit us to extrapolate sample characteristics to a total population of interest within known confidence intervals.

In a controlled experiment, on the other hand, an explicit attempt is made to apply stimuli to a sample of households in order to observe induced changes in behavior, and then to relate the results to the effects of applying similar stimuli to the total population on a non-experimental basis. The position taken by economists was that the prototype design model--with its comparison of independent "snapshots" of the population to measure experimental effects--was insufficient for handling the complex hypotheses to be tested. To them, experimentation meant exerting control over variations in program parameters and stratification characteristics--to permit estimation of structural relationships with random residuals. The emphasis of survey statisticians on random draws from populations and estimation of population means was of lesser importance.

The lack of correspondence between responses observed in experiments and program effects of direct policy interest comes from a number of sources.

First, the program to be ultimately considered for implementation is not known at the time experimentation begins, and is unlikely to correspond exactly to any of the treatments being experimented with. Thus, it may be necessary to extend experimental results to programs having similar but not identical characteristics.

Second, certain options considered for program implementation may not be viable subjects of experimentation. Since participation in social experiments is a voluntary process, the effects of policy options which leave some individuals worse off than under existing programs cannot be observed directly. Thus experimental results sometimes must be extrapolated beyond the scope of the tested programs.

Third, the results of an experiment depend both on the experimental program structure and on the environment faced by the control group. This background environment may differ from what is to prevail at the time of program implementation; thus it is important to standardize the environment of the control group where possible, and to understand its effects. Control of the background environment has proved to be one of the major problems in the social experiments. During the New Jersey experiment, for example, there were two major changes in the welfare system not only affecting the control group but also providing benefits more generous than those paid by some of the experimental treatments.

Fourth, certain features of an implemented program are virtually impossible to replicate or to observe in an experimental environment. The New Jersey experiment provided payments for only three years, while an implemented program would be of permanent duration. Some implemented programs--such as the transitional employment associated with Supported Work--would be similar in duration to their experimental counterparts, but their long-run effects may only be apparent after the results of the evaluation are required. Full scale implementation of a program may lead to different effects than those of a sample blown up to the full eligible population. For example, if the NIT had a major effect on the labor supply of low income households, it would have an impact on labor markets and wage rates unobservable in a small sample experiment.

Finally, there is the issue recognized from the very start of the experiments but not fully confronted: the possibility that individuals who are in an experimental setting may react differently than they would under normal circumstances.

Thus, the experiments do not provide direct answers to policy questions, but must be supplemented by nonexperimental analytic techniques. An integral part of the experimental process must be the provision of the information necessary for such analyses.

F. CONCLUSIONS

Prior to the advent of the social experiments, economists and other social scientists developed quantitative techniques for testing hypotheses with nonexperimental data. They developed methods of applied statistical inference which required prior acceptance of structural specifications. Econometrics became a well-developed form of applied statistics, and economists have long turned to their own profession for guidance in this area. The contribution of statisticians, on the other hand, could have been in the areas of sampling methodology and experimental design. While statisticians were consulted at various stages of the social experiments and made some valuable contributions, economists played a dominant role in design decisions.

During its early days, social experimentation was viewed as a technological revolution, and perhaps too much was expected of it. The social experiments are flawed in what they can do--not only because of errors in execution by economists and others--but also because creation of the appropriate experimental environment may be conceptually impossible. To be used correctly, therefore, experimentation must be viewed as an augmentation to existing methods of program evaluation rather than as a radical departure. Social experimentation exists today as a viable methodological tool because of economists and policy makers willing to listen to them; in the process it has acquired both the strengths and the weaknesses of their methodological perspective.

FOOTNOTES

1. The author wishes to thank David N. Kershaw and Cheri T. Marshall for their comments and contributions to the content of this paper. The views expressed here are the sole responsibility of the author.

2. For a review of the origins and design of the New Jersey Experiment, see Kershaw and Fair (5). See Rossi and Layall (8) for a major external critique of the experiment. Rossi is the leading critic of the experiment from a sociological perspective.

3. Ross (7).

4. Scranton, Pennsylvania was added to the original set of New Jersey cities after the New Jersey sample proved to be predominantly black and Puerto Rican.

5. See Hausman and Wise (3).

6. Portions of this and the following section draw freely from Metcalf (6).

7. Some economists have argued that experimentation with "extreme" treatments is useful in ways analagous to the use of extreme dosages in medical experiments.

8. See Conlisk (1) and Conlisk and Watts (2) for a discussion of the conditions under which orthogonality is desirable.

9. See Conlisk and Watts (2), Metcalf (6), and Rossi and Lyall (8) for detailed discussions of the Conlisk-Watts model.

10. In particular, see the discussion of truncated sampling frames in Section B above.

11. See Hollister and Metcalf (4) for a discussion of this issue.

REFERENCES

1. Conlisk, J., "When Collinearity is Desirable. Western Economic Journal 9 (1971): 393-407.

2. Conlisk, J. and Watts, H., "A Model for Optimizing Experimental Design for Estimating Response Surfaces." 1969 Proceedings of the Social Statistics Section, American Statistical Assn., pp 150-156.

3. Hausman, J., and Wise, D., "Social Experimentation, Truncated Distributions, and Efficient Estimation" in Followup Studies Using Data Generated by the New Jersey Negative Income Tax Experiment, Mathematica Policy Research, March 1976.

4. Hollister, R. and Metcalf, C.E., 1977. "The Labor Supply Response of the Family," in The New Jersey Income-Maintenance Experiment, vol. II: Labor Supply Responses. ed. by Harold W. Watts and Albert Rees, New York: Academic Press.

5. Kershaw, D. N. and Fair, J. 1976. The New Jersey Income-Maintenance Experiment, vol. I: Operation, Surveys, and Administration. New York: Academic Press.

6. Metcalf, C. E., "Sample Design and the Use of Experimental Data," in The New Jersey Income-Maintenance Experiment, vol. III. Expenditures, Health, and Social Behavior; and the Quality of the Evidence. Edited by Harold W. Watts and Albert Rees, New York: Academic Press (forthcoming).

7. Ross, H., "A Proposal for a Demonstration of New Techniques in Income Maintenance," Memorandum, December 1966, Data Center Archives, Institute for Research on Poverty, University of Wisconsin, Madison.

8. Rossi, P. H. and Lyall, K. C. 1976. Reforming Public Welfare: A Critique of the Negative Income Tax Experiment. New York: Russell Sage Foundation.