NBER WORKING PAPER SERIES

TECHNICAL PROBLEMS IN SOCIAL EXPERIMENTATION: COST VERSUS EASE OF ANALYSIS

Jerry A. Hausman

David A. Wise

Working Paper No. 1061

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge MA 02138

January 1983

Any opinions expressed are those of the authors and not those of the National Bureau of Economic Research.

NBER Working Paper #1061
January 1983

Technical Problems in Social Experimentation: Cost Versus Ease of Analysis

ABSTRACT

The goal of the paper is to set forth general guidelines that we believe would enhance the usefulness of future social experiments and to suggest ways of correcting for inherent limitations of them. Although the major motivation for an experiment is to overcome the inherent limitations of structural econometric models, in many instances the experimental designs have subverted this motivation. The primary advantages of randomized controlled experiments were often lost. The major complication for the analysis of the experiments was induced by an endogenous sample selection and treatment assignment procedure that selected the experimental participants and assigned them to control versus treatment groups partly on the basis of the variable whose response the experiments were intended to measure. We propose that to overcome these difficulties, the goal of an experimental design should be as nearly as possible to allow analysis based on a simple analysis of variance model. Although complexities attendant to endogenous stratification can be avoided, there are inherent limitations of the experiments that cannot. Two major ones are self-determination of participation and self-selection out, through attrition. But these problems, we believe, can be corrected for with relative ease if endogenous stratification is eliminated. Finally, we propose that as a guiding principle, the experiments should have as a first priority the precise estimation of a single or a small number of treatment effects.

Jerry A. Hausman
Economics Department
MIT
Cambridge, MA 02139
(617) 253-3644

David A. Wise
Harvard University
J.F.K. School of Government
Cambridge, MA 02138
(617) 495-1178

TECHNICAL PROBLEMS IN SOCIAL EXPERIMENTATION: COST VERSUS EASE OF ANALYSIS

by

Jerry A. Hausman and David A. Wise

Over the past decade, a major portion of empirical economic research has been based on what have come to be known as social experiments. Primary examples include a series of income maintenance experiments, a housing allowance demand experiment, several electricity pricing experiments, and a health insurance experiment. Much of our discussion in this paper is motivated by the income maintenance experiments but it also draws from our experience with the housing allowance and electricity experiments as well.

The goal of the paper is to set forth general guidelines that we believe would enhance the usefulness of future social experiments and to suggest ways of correcting for inherent limitations of them. Our conclusion and results may be summarized briefly.

Although the major motivation for an experiment is to overcome the inherent limitations of structural econometric models, in many instances the experimental designs have subverted this motivation. The primary advantages of randomized controlled experiments were often lost. In particular, it was in large measure impossible to estimate an experimental effect using straightforward analysis of variance methods, as a standard experimental design would suggest. Rather, a careful analysis of the results often required complicated structural models based on strong model specification assumptions, the necessity for which an experiment should be designed to obviate. Section I provides a simple explanation of this goal and is intended to motivate the remainder of the paper.

The major complication for the analysis of the experiments was induced by an endogenous sample selection and treatment assignment procedure that selected the experimental participants and assigned them to control versus treatment groups partly on the basis of an outcome variable the change in which, the experiments were intended to measure. To overcome at the time of analysis of the experimental results the complications caused by the endogenous sample selection and treatment assignment required rather complex statistical techniques and detracted greatly from the simplicity we believe should be a goal of experimental designs.

We propose that to overcome these difficulties, an experimental design should as nearly as possible allow analysis based on a simple analysis of variance model. This would mean that sample selection and treatment assignment should be based on randomization and that stratification on response variables should be avoided.

Although complexities attendant to endogenous stratification can be avoided, there are inherent limitations of the experiments that cannot. Two major ones are self-determination of participation and self-selection out, through attrition. But these problems, we believe, can be corrected for with relative ease if endogenous stratification is eliminated.

Finally, we propose that as a guiding principle, the experiments should have as a first priority the precise estimation of a single or a small number of treatment effects. The experiments to date have in general been hampered by a large number of treatments together with small sample sizes so that no single treatment could be estimated accurately.

Following the motivation in Section I, we have elaborated in Section II these several general guidelines that we believe would enhance the effectiveness of future experiments. The problem of endogenous stratification and a way of avoiding it are set forth in Section III. A method of correcting for the inherent self-selection problems of social experiments is suggested in Section IV.

I. Unbiased Estimates, Structural Models, and Randomization

To obtain unbiased estimates is the major motivation for a large portion of econometric theory and for the application of econometric techniques in empirical analysis. Econometricians generally have in mind a model of the form

$$Y = f(X, \varepsilon) \tag{1}$$

where X represents measured and ε unmeasured determinants of Y. The goal is to estimate the effects of the elements of X on Y. A common specification of f in (1) is

$$Y = X\beta + \varepsilon , \tag{2}$$

where β is a vector of parameters to be estimated, with each element of β measuring the effect on Y of a unit change in the corresponding element of X.

The guiding principle for econometricians is that simple estimation techniques (e.g., least squares) will yield unbiased estimates of β if X is uncorrelated with ε. Unbiased is understood to mean and is indeed defined to mean an unbiased estimate of the "causal" effect of X on Y, the understood definition of β in much, but not all, of econometric analysis. But although the principle is demonstrably true in theory, it is often difficult to approximate in practice and its existence impossible to verify without reservation. Nonetheless, the goal remains.

To move toward it, econometricians use two general modes of reasoning. One is economic theory that restricts the functional form of f, although usually only within broad bounds. The other is statistical theory, that in large part prescribes methods to correct for correlation between X and ε, and thus to obtain unbiased estimates of β. The combination of economic and statistical theory often leads--at least in the abstract--to specification and estimation of structural models. Structural models can be thought of as those in which the parameters have a causal interpretation, and with the concomitant property that if unbiased estimates of them are obtained they also could be given a causal interpretation. But although theoretical prescription of models and their empirical estimation can restrict the form of f, they can do so only within limits. The estimates must be interpreted within the constraints implicit in the assumptions that underlie them. In particular, it is usually not possible to know for sure that X is uncorrelated with ε, or if not, that corrections have been made for correlations that exist.

A response to this dilemma is to choose selected values of X in such a way that they are by design uncorrelated with other determinants of Y, and thus allowing unbiased estimation of the corresponding values of β. The technique is randomization and it is most often employed within the context of a randomized controlled experiment. For purposes of exposition, we shall henceforth use as an example estimation of the effects of income maintenance plans--taxes and guarantees--on earnings.

Suppose that the plan is T, called the treatment, and that earnings depend on T, other measured variables X, and on unmeasured determinants ε according to

$$Y = \beta_1 T + f(X, \varepsilon) \ . \tag{3}$$

If individuals (more often families) are chosen at random from the population and assigned values of T, in large samples T will be uncorrelated with ε and with X as well. Then simple least squares analysis of variance estimation of the model

$$Y = \beta_1 T + \eta \tag{4}$$

where η is equal to f and treated as a disturbance term in this model, will yield unbiased estimates of β.

The primary motivation for this approach is to circumvent the uncertainties inherent in the assumptions of structural econometric models, by constructing T in such a way that it is uncorrelated with other determinants of Y and thus by construction assuring unbiased estimation of β_1.

We have set forth these possibly oversimplified ideas to serve as background and motivation for our subsequent discussion. In particular, it is important to keep in mind the motivation for randomized controlled experiments. Although in the large social experiments, we believe it is impossible to create the theoretical paradigm of such an experiment, we believe that the paradigm should serve as a guide to their designs as well as to the analysis of their results--much as the theoretical goal of X's uncorrelated with error terms serves as a guide to empirical analysis based on non-experimental data. We shall argue, for example, that the use of complex structural models to analyze the data from social experiments, or experimental designs that require such models or depend in large part on structural model assumptions, are often in contradiction to the primary motivation for the experiments and thus subvert their intent; they are often inconsistent with the raison d'être of experiments. We will elaborate more on this and other general propositions in the next section.

II. General Goals and Guiding Propositions

With the powerful advantage of hindsight, and we hope aided by our part in the analysis of social experiments to date, we shall set forth several propositions that we believe will enhance the value of future experiments. To do this, we will explain what we believe to be the major inherent limitations of such experiments. The primary ones are self-determination of experimental participation and self-determination of withdrawal from the experiment. These we believe can be corrected for, and some suggestions for doing so are contained in the following sections. There are other design characteristics of the experiments to date that we believe unnecessarily complicated their analysis, and in particular made it much more difficult to correct for the inherent limitations of them. The primary design feature of this type is stratification on endogenous variables. We will address this question first, then turn to a discussion of inherent limitations, and then address other principles that we believe should guide future experimental designs.

A. Stratification on Endogenous Variables

As described in the previous section, the reason for an experiment is, by randomization, to *eliminate* correlation between the treatment variable and other determinants of the response variable that is under study. In each of the income maintenance experiments, however, the experimental sample was selected in part on the basis of the dependent variable and the assignment to treatment versus control group was based in part on the dependent variable as well. In general, the group eligible for selection--based on family status, race, age of family head, etc.--was stratified on the basis of income (and other variables) and persons selected from within the strata. In the New Jersey experiment, persons with incomes greater than 1.5 times the poverty level were excluded altogether. In the other experiments, the stratification on income was less complete, but as a result a bit more complicated. Assignment to control versus treatment group was also based in part on income. Whether the outcome of interest is income or hours worked, which is a component of income, such a procedure *induces* correlation between right-hand variables, including the treatment effect, and unmeasured determinants of income. Thus it is not straightforward to obtain unbiased estimates of treatment effects using simple analysis of variance or covariance techniques.

Theoretically, a very elaborate analysis of variance procedure that allowed for estimation of separate treatment effects within each strata would yield unbiased estimates. But because the strata were so numerous and the treatments so many and the sample sizes relatively small, this method of analysis was impractical because reasonably precise estimates of treatment effects could not be obtained. Thus to correct for endogenous stratification and treatment assignment required rather complicated models (Hausman and Wise [1977], [1979], and [1980]).

Analysis of experimental results based on such techniques has at least two major shortcomings. First, it is relatively complicated--requiring non-linear maximum likelihood estimation for example. This is a shortcoming in itself, but seems especially troublesome in the context of an experiment one of whose major advantages presumably is simplicity. Second, and more important, it necessitates the imposition of functional form constraints. The models proposed by Hausman and Wise are generally structural in spirit, and in particular require distributional assumptions against which the results may not be robust. To correct for endogenous stratification, for example, requires analysis based on truncated distributions in which the distribution assumed is necessarily a key component. Since the primary advantage of an experiment presumably is to lessen or avoid the necessity for such assumptions, it seems contradictory to design experiments whose effects cannot be evaluated accurately without them.

The elimination of stratification on endogenous variables would avoid this source of complication. The most straightforward procedure would be to randomly select an experimental group from the population and randomly assign those selected to control or treatment status, without consideration of income or other <u>endogenous</u> variables. There seem to be two major objections to such a procedure: cost and political feasibility. Indeed the two are not unrelated. Most seriously considered income support programs are intended to guarantee a minimum income to families who would otherwise have relatively low incomes. And presumably it is primarily this group whose labor supply and earnings would be affected by the plan. Nonetheless, it has been difficult to obtain funds for experimental programs that guaranteed support for higher income families, even though under most plans payments to this group would be small, since their earnings would be unlikely to fall below the "breakeven" point at which payments are zero. In addition, if it is important to obtain a "good" estimate of the effect of the program on low income families, then it is necessary to have a large enough number of low income families to do so. Of course a large random sample from the population would also provide a large number of low income families. But larger sample sizes of course increase the cost of the experiment.

We do not present numbers on the marginal cost of an additional experimental family. Preliminary investigation, however, suggests that it is small relative to the fixed costs of running an experiment. Suppose that for whatever reason, it is not feasible to select a random sample from the population. We propose in this case that the sample be as random as possible. That is, randomly select persons with incomes below a given level, without endogenous stratification within this group. But what should be the measure of income that determines eligibility?

We have proposed in Section III--after a more detailed description of the endogenous stratification problem--a method for selecting the experimental group, based on predicted income, in such a way that the stratification is not endogenous.

B. Inherent Limitations on Random Sample Selection

We have argued that endogenous stratification procedures unduly complicate the analysis of experimental results and that procedures that avoid such stratification would be preferable. Nonetheless, there are inherent limitations on randomization in social experiments. It is surely impossible to attain the theoretical paradigm of a randomized-controlled experiment. There are at least two major reasons for this, both involving individual self-selection.

One is that persons cannot in general be made to participate in an experiment if selected by a random procedure. Some of those randomly selected will participate while others will not. If the individual participation decision is related to the effect that the treatment would have on individuals, then the estimated treatment effect will be a biased estimate of the effect to be expected if the treatment were instituted as a program applying to the entire population.

The 1954 Salk vaccine experiment provides a good example of this effect. There were two primary versions of the experimental design. In the "placebo control" areas, children who agreed to be inoculated (or, more accurately, whose parents agreed to the inoculation) were randomly assigned to the vaccine group or to the placebo group. In the "observed control" area, second grade children who agreed to inoculation received the vaccine, while first and third graders served as the control group. Selected results are shown in table 1.

Children in the placebo control areas who were not inoculated contracted polio at a rate of 54 per 100,000. The comparable figure for children who participated in the experiment was 81, the rate for those who participated and received the placebo. Similarly in the observed control areas, grade 2 children who were not inoculated had a substantially lower rate: (53), than the rate for the control group (61). Thus apparently children who were more likely to contract polio and thus more likely to be helped by the vaccine, were more likely to participate in the experiment. This tends to exaggerate the effect of the vaccine. For example, one might conclude on the basis of the vaccinated and control groups in the observed control areas that the vaccine reduced the rate from 61 to 34. But apparently the rate for all children would have been less than 61 without the vaccine. It is of course apparent from these data that the vaccine was effective, regardless of this uncertainty about the magnitude of the effect. But if the effect had been less clear, this self-determination of participation could have led to considerable uncertainty about desirability of universal inoculation.

A similar effect was apparent in the recent housing allowance demand experiment. Because of the nature of the primary experimental allowance, many families could benefit under the allowance plan only if they were willing to move. It seems apparent from subsequent analysis that of low income renters who were asked to participate in the experiment, those who were less adverse to moving were more likely to participate in the experiment. (See Venti and Wise [1982].) Thus the estimated experimental effect tended to exaggerate the increase in rent that would be induced by the allowance were it applied to all low income renters.

We have suggested in Section IV a procedure that we believe could be used to correct for this potential bias, assuming that the self-selection cannot be avoided.

The other form of self-selection is attrition from the experimental sample, once a sample has been selected. Again, the problem is that determinants of dropping out may be related to the experimental response that would otherwise be observed. For example, persons who are not affected by the treatment, possibly because they have high incomes for example, may be more likely to drop out than those who are affected and thus receive higher payments. This is the problem addressed by Hausman and Wise [1979].

If the experimental design is not complicated by endogenous stratification and assignment, then correction for self-determination of participation and attrition would be relatively simple. Indeed correction for both simultaneously is quite feasible and this is the approach taken in Section IV. Such a correction, however, is much more complicated if the experimental design is also complicated by endogenous stratification and assignment. This reinforces the proposal that such stratification be avoided in favor of random sampling. Then analysis of experimental results can address complications that are unavoidable without having to devote extraordinary effort to correct for complications induced by the experimental design.

C. Additional Concerns

A characteristic of experiments to date has been a rather large number of treatments. The income maintenance experiments, for example, entailed several treatments defined by different combinations of income guarantee levels and tax rates. In none of the experiments, however, were the sample sizes large enough to obtain precise estimates of the effects of any particular treatment. Thus analysts generally resorted to estimation of a single effect that did not distinguish the various treatments, or they assumed a structural model that allowed interpolation across individuals assigned to different treatments. The more the latter procedure was followed, the less consistent the analysis was with the motivation for an experiment. That is, it subverted the major goal of using random selection and treatment assignment to circumvent the inherent limitations of hypothesized structural models.

Thus it seems to us that priorities should be ordered in such a way that the primary goals of an experiment are met first. The first goal we propose should be the estimation of an experimental effect for _a_ treatment. Then additional treatments should be added only if each additional one can also be estimated with precision. The proposition is that precise estimation of the effect of single treatment or the effects of a few treatments is to be preferred to imprecise estimates of many.

This we propose should be done in such a way that simple analysis of covariance estimates of treatment effects may be obtained, subject to the limitations on randomization discussed above and detailed more fully below. Thus we would propose an evaluation model of the form

$$Y = \alpha_1 T_1 + \alpha_2 T_2 \cdot \cdot \cdot \cdot + \alpha_k T_k + X\beta + \varepsilon$$

where the α_k are treatment effects. We propose an analysis of covariance model because our research (Hausman and Wise [1979]) has suggested that the use of exogenous control variables, represented by X, reduces the effect of attrition on estimated experimental effects and we presume that it would be likely also to reduce the effect of self-determination of participation.

The reader will note the absence of a structural parameterization that attempts, for example, to describe income and substitution effects. This is because we believe that simple precise estimates of a few effects will be more readily understood by most observers and will thus carry more weight in the decision-making process. In addition, if for policy purposes, it is desirable to estimate the effects of possible programs not described by treatments, then interpolations can be made between estimated treatment effects. If the experimental treatments are at the bounds of possible programs, then of course this is easier. Although it can be argued that structural models are necessary to make interpolations, we believe that for almost any situation we can think of, the simplicity of say linear interpolations far outweigh the possible advantages of interpolations based on a structural model. At the same time, it maintains the spirit of an experiment.

If the experiment is to inform the policy making process, we believe that a single number that can be supported can be more confidently relied on than more complex analysis. That the labor supply effect of a known treatment is 16 percent and not 2 percent, for example, is we believe much more important than whether the effect of a plan close to the treatment is 16 percent or 17 percent.

This is not to say that experimental data should not be used to estimate structural econometric models. These data can of course be used like other survey data for this purpose. But the experiment should be thought of in the first instance as a way to obtain accurate estimates of the effects of particular programs. Structural models with parameters estimated on survey data could also be used to make such estimates. (Presumably this would be done to a considerable extent before an experiment were undertaken, if for no other reason than simply to help to inform the choice of experimental treatment or treatments.) In this sense, the experiment could be thought of as checking the accuracy of predictions based on analysis of survey data. That is, the experiments should be designed to provide a selected number of points "on" the response surface, defined for example by tax rate and guarantee levels. It is rather straightforward to check for example the degree to which alternative structural models fit these "known" points on the response surface. In short, an experiment should be used to avoid the inherent limitations of structural models in providing accurate estimates of the effects of specified programs. Their major advantage should not be lost sight of in an effort to estimate models that will predict the result of any plan. A lack of confidence in such estimates is the motivation for the experiments. To use the experimental data only to provide more such estimates, or to set up the experiments in such a way that only such estimates are possible, is to travel to Rome to buy canned peas.

III. Endogenous Sampling and Stratification

As discussed in the introduction above, a major feature of classical experimental design is that it leads to a simple analysis of variance (ANOVA) model that minimizes the number of maintained assumptions implicit in the interpretation of parameter estimates. That is, the analysis is "model free" in two important aspects: (1) In the simplest cases a main effects ANOVA specification is adequate. Questions about the need to include further right-hand variables--as in much of econometric and statistical analysis--for example, do not arise. Correct randomization assures that disturbance terms have expectation equal to zero. Also, questions of functional form are absent because each experimental treatment effect is measured by a parameter. (2) Distributional assumptions are kept to a minimum in estimation. While distributions of test statistics are certainly used in inference, asymptotic theory may provide a reasonably good approximation in many cases. Classical experimental design together with ANOVA offer the opportunity either to eliminate or to decrease greatly a major problem that arises in econometric studies based on observational, i.e., non-experimental data.[1]

Yet in many of the social experiments the classical approach has not been followed. Given a limited experimental budget and a "target population," the designers of the experiments, in concentrating sample selection on that part of the population most likely to be affected by the treatment policy, induced endogenous sample selection and treatment assignment. The presence of endogenous sampling complicates the analysis of the experiment greatly and thus limits our ability to treat other problems which arise, in particular, sample self-selection and attrition. And possibly as important, it typically forces the analyst to maintain distributional assumptions about the random variables under study. These distributional assumptions are not innocuous even in large samples. Significant empirical departures from these assumptions may lead to large biases in estimation of experimental effects (e.g., Goldberger 1980). Most importantly, if the endogenous sampling is ignored in the analysis, extremely large biases may result in estimated experimental effects. In this section we will present three examples of endogenous sampling as well as techniques developed to eliminate the problems that it creates. We then propose an alternative approach which attempts to choose selectively from the target population without inducing endogenous sample selection.

The problems associated with endogenous sampling occur because a pre-experimental endogenous variable is used in sample selection and in treatment assignment. The effect on the estimated treatment effect arises because of correlation between unmeasured determinants of the response variable in the experimental and pre-experimental periods. These time effects have often been ignored in the experimental designs.[2] We shall illustrate the problem within the context of an ANOVA framework, which when generalized to a random effects specification allows for serial correlation. We consider a single period experiment with one period of pre-experimental data.

$$
\begin{aligned}
&Y_{it} = u_t + \beta_j T_{jt} + \mu_i + \eta_{it} ; \\
&t = 1,2; \quad j = 1, \ . \ . \ . \ , J . \\
(5) \qquad &E\mu_i = E\eta_{it} = 0; \ V(\mu_i) = \sigma_\mu^2 ; \\
&V(\eta_{it}) = \sigma_\eta^2; \qquad \rho = \frac{\sigma_\mu^2}{\sigma_\eta^2 + \sigma_\mu^2}
\end{aligned}
$$

We have decomposed the disturbance term into a permanent individual component μ_i and another component η_{it} assumed independent across time periods.[3] The indicator variable T_{jt} is 1 if the individual is receiving the experimental treatment j in period t and zero otherwise. Time effects are absorbed into the constant terms u_t. The importance of the individual component μ_i is given by the correlation ρ between the disturbance term in the two time periods. Such correlations often exceed .5 in econometric studies.

Now suppose that the expected cost of an experimental treatment varies across individuals and treatments as a function of Y_{i1}. Designers of experiments have for this reason used Y_{i1} in sample selection and in treatment assignment. Because of the presence of μ_i in both periods the endogenous sampling and treatment assignment based on pre-experimental data carries over to the experimental period as well. A simple example will help to make the point clear. Suppose we have two experimental treatments called generous (G) and not-generous (NG). The G treatment is expected to cost more for "high Y" individuals because of an expected percentage reduction in work effort. Therefore, the designer forms two groups of individuals based on Y_{i1}. Low Y_1 individuals are assigned either the G plan or control status; the high Y_1 individuals receive either the NG plan or control status. But when we use ANOVA to analyze the experimental results we see from equation (5) that $E(\mu_i|T_{jt}) \neq 0$. Thus, our estimates are biased for the population since we have not accounted for the presence of individual effects that persist over time. Since it is unlikely in most economic and social experiments that ρ is near zero, substantial biases may arise from endogenous sample designs.

We shall now consider three experimental designs in which endogenous sampling was used. (i) In the New Jersey Negative Income Tax experiment any individual whose pre-experimental income exceeded 1.5 times the government set poverty limit was excluded from the sample. This sample truncation was used because the major effect of an NIT program was expected to be on low income individuals and families. A simple rule was thus used to make the sample resemble the target population. Suppose a model like equation (5) is used to analyze the effects on hours worked. Suppose also that individuals' earnings are low in period one either because they have low μ or because η_1 is negative even though μ is positive. Low μ people with positive η_1 have been excluded from the sample. The analyst must maintain the assumption that the effect on hours worked for the sample combination of low μ and high μ people (with negative η) will represent the total population response. This assumption appears unlikely to hold true because we might well expect the behavioral response to differ among the low μ and high μ people. In other words, if we were to change the sample truncation point from 1.5 times the poverty limit to another level, the estimated experimental effect would be likely to change as well.

(ii) In the Connecticut Time of Day Electricity Demonstration (1977), the sample was grouped into quintiles on the basis of electricity usage in the year prior to the demonstration. Then households in the upper quintiles were disproportionately sampled since the electric utility correctly thought that their reaction to the introduction of time of day electricity rates would have the largest effects on system revenues.

(iii) In the Seattle-Denver Income Maintenance Experiment, (SIME-DIME), the Conlisk-Watts framework was used for treatment assignment. It allowed the expected cost of an experimental treatment c_j for treatment T_j to vary with "normal income" which in practice was very closely related to pre-experimental income. Consider the Conlisk-Watts framework in the regression form

$$(6) \quad \begin{aligned} &Y = X\beta + \varepsilon \ ; \\ &X_j = (0, \ . \ . \ . \ , 0, 1, 0, \ . \ . \ . \ 0); \ j = 1, J \ ; \\ &E\varepsilon = 0 \ ; \\ &V(\varepsilon) = \sigma^2 I \ . \end{aligned}$$

Here X_1 denotes the control observations and the $j = 2, \ . \ . \ . \ J$ denote the $J - 1$ experimental treatments and normal income classifications. The Conlisk-Watts design uses as an optimization criterion the minimization of the variance of linear functions $P\hat{\beta}$ of the estimated coefficients, subject to a budget constraint. We want to choose n_j, $j = 1, J$ (the number of individuals in a given row of the design matrix) in an optimal manner. Let $D = P'P$. The complete problem is an integer programming problem with a convex objective function subject to linear constraints.

$$\text{(7)}\qquad \begin{aligned} \min q(n_1, \ldots, n_m) &= \operatorname{tr}[D \sum_{j=1}^{J} n_j x_j' x_j)^{-1}] , \\ n_j &\geq 0 \text{ for all } j . \end{aligned}$$

For large $N = \Sigma n_j$ a suitable approximation is to treat the n_j as continuous and to round off the results to the nearest integer. To estimate the experimental effects in each class via the contrasts, $\hat{\beta}_j - \hat{\beta}_1$, the appropriate P matrix is an $(m - 1) \times m$ matrix with the first column -1's and each of the remaining columns with all zeroes and a single 1. Thus $P_j = [-1, 0, \ldots, 0, 1, \ldots, 0]$. We solve equation (7) to find

$$\text{(8)}\qquad \begin{aligned} n_1 &= C \frac{((J - 1)/c_1)^{\frac{1}{2}}}{E} , \\ n_j &= C(c_j^{-\frac{1}{2}} E^{-1}) , \\ E &= [(J - 1)c_1 + \sum_{j=2}^{J} c_J]^{\frac{1}{2}} . \end{aligned}$$

The optimal design thus increases the probability of inclusion in the sample for low c_j individuals. But since c_j is a function of pre-experimental income we see that $E(\mu_i | X_j) \neq 0$ which will lead to bias in the estimation of experimental effects.

We do not want to give the erroneous impression that endogenous sampling destroys the possibility of experimental analysis. In fact, we have written several papers addressing the problem, Hausman-Wise

[1976, 1977, 1980, 1981]. And endogenous sampling can reduce the cost of an experiment considerably.[4] But we emphasize the model functional form and distributional assumptions that endogenous sampling requires.

To illustrate the nature of these assumptions, we consider again the three examples and for each we discuss possible model specifications.

(1) Sample truncation: In Hausman and Wise [1976, 1977], models to correct for sample truncation are developed. The approach taken assumes that the earnings conditional on personal attributes are distributed log-normal. A two period model is necessary since sample truncation was performed on the pre-experimental data. But since the correlation of the disturbances across years (ρ in equation (5)) is not zero, truncation on pre-experimental data will affect the analysis of the experimental results. Therefore, we define a model of the form

$$(9) \qquad y_{it} = Z_{it}\gamma + \varepsilon_{it} \; ; \qquad t = 1, 2 \; ; \qquad \varepsilon_{it} = \mu_i + \eta_{it} \; .$$

with the usual stochastic assumptions. We assume that $f(y_{i1}, y_{i2} | Z_{i1}, Z_{i2})$ is bivariate normal. The Z_{it}'s include experimental treatments as well as individual characteristics. Then the likelihood can be written

$$(10) \qquad L = \prod_{i=1}^{N} f(y_{1i}, y_{2i}) = \prod_{i=1}^{N} \frac{\tilde{\phi}(y_{1i}, y_{2i})}{\Phi[(L_i - Z_{1i}\beta)/\sigma]} \; .$$

where $\tilde{\phi}$ is the bivariate normal density and Φ is the univariate normal distribution. For the New Jersey NIT experiment we estimate $\hat{\rho} = .85$ which demonstrates the potential importance of correcting for truncation. The log normal is a convenient distribution which leads to a likelihood function that is quite tractable using modern computers. Still, if the choice of log normal is not correct, it represents a specification error.

An even more difficult problem arises if we want to analyze hours rather than earnings. Since truncation takes place on earnings we must analyze hours and wages jointly and the four-equation model that results leads to a likelihood function that is considerably more complicated than equation (10). (See Hausman and Wise (1976, p.432).) Furthermore, given the identity between earnings and the product of wages and hours, we must now assume that both wages and hours are distributed log normally. Almost no other assumptions lead to a tractable likelihood function even though some evidence exists that hours might be better represented by a conditional normal distribution.[5] And lastly, because of the complications induced in the likelihood function by truncation, our ability to handle other problems, like sample attrition or taxation, are limited. Thus the analysis has been greatly complicated by what seems to be a reasonable design criterion, concentrating on the target population of the proposed policy.

(2) Stratification on the endogenous variables: To keep the analysis simple we here assume that income has been grouped into two intervals, even though in the Gary NIT experiment as well as the Connecticut TOD demonstration quintiles were used. Assume that below some level L an unknown proportion of a random sample of the population is sampled, P_1, and above L, a proportion P_2.[6] Then the density function is

$$(11) \qquad h(y) = \begin{cases} \dfrac{P_1 \cdot f(y)}{P_1 . Pr[y \leq L] + P_2 \cdot Pr[y > L]} , & \text{if } y \leq L \\[2ex] \dfrac{P_2 \cdot f(y)}{P_1 \cdot Pr[y \leq L] + P_2 \cdot Pr[y > L]} , & \text{if } y > L , \end{cases}$$

where f is the normal density function $N(Z\beta, \sigma^2)$. Only the ratio $P = P_2/P_1$ can be identified. Therefore, we divide through the expressions in equation (11) by P_1. Again using the normality assumption for y and assuming N_1 persons with $y \leq L$ and N_2 with $y > 1$ the log likelihood function is

$$L = \sum_{i=1}^{N_1} \ln f(y_i) - \sum_{i=1}^{N_1} \ln(\Phi_i + P(1 - \Phi_i)) +$$

$$(12) \qquad \sum_{i=1}^{N_2} \ln P + \sum_{i=1}^{N_2} \ln f(y_i) - \sum_{i=1}^{N_2} \ln (\Phi_i + P(1 - \Phi_i))$$

$$\sum_{i=1}^{N} \ln f(y_i) - \sum_{i=1}^{N} \ln (P + (1 - P)\Phi_i) + N_2 \ln P .$$

where $\Phi_1 = \Phi[(L - Zi\beta]$. Again, a maintained distributional assumption is necessary and a rather complicated maximum likelihood problem is presented. Furthermore, when we want to do a two period analysis or consider other problems, our ability to do so is limited by the rapidly increasing complications induced by the stratification on the endogenous variable.

(3) Treatment assignment using an endogenous variable: Our last example is the SIME-DIME NIT experimental design. Here seven income

intervals, called "E-levels,' were used to define rows in the Conlisk-Watts design framework of equations (6)-(18). The costs c_j were then derived as a function of E-level. The expected cost of a treatment was presumed to rise with E-level because it was assumed that tax revenues would decline and that NIT payments would increase. The result was that no one in the highest E-level interval was assigned treatment status; all were assigned to be controls where, of course, the cost does not grow with E-level. Furthermore, in general persons with higher E-levels were more likely to be assigned to experimental treatments with more generous support levels. Thus, treatment assignment was based on an endogenous variable, pre-experimental income, which was highly correlated with the response variable during the experiment.

Treatment assignment using endogenous variables does not in theory prevent the use of ANOVA in the analysis phase of an experiment. What is needed, however, is an elaborate specification allowing a separate β in equation (5) for each E-level and treatment or control assignment. But, in the SIME-DIME experiment for example, including manpower treatments, there would be J = 59 columns in the X matrix. In fact, if full ANOVA were done without deleting higher order interactions as did the design model, we would have J exceeding 200. Thus, even for the comparatively large sample sizes as in the SIME-DIME we cannot hope to obtain precise estimates of experimental effects. And when other factors such as race and city are added to the analysis, full ANOVA specifications with many fewer parameters than the experimental design requires. One approach is to enter E-level as a right-hand side variable in linear form. But we immediately lose the model free aspect of ANOVA since correctness of functional form becomes an issue.

In fact, a linear specification of E-level is not totally appropriate since it does not remove all correlation between the treatment variable and the stochastic disturbance.

Again, a model of treatment assignment can be constructed, which Hausman-Wise [1980] specify. But since treatment assignment is a zero-one outcome, a probit model (or logit model) is required along with the necessary distributional assumptions. An additional complication arises here because we must also specify the partly unknown model of treatment assignment correctly.[7] Thus, both distributional assumptions and functional form assumptions are required for model estimation. The resulting likelihood function used in estimation is even more complicated than equations (10) and (12). And as emphasized above, additional complications like sample attrition are almost impossible to treat jointly with the sample assignment issues.

A very simple solution exists to these design and analysis problems. Randomize over pre-experimental income. Then problems of endogenous assignment or stratification do not occur, so that ANOVA specifications again are appropriate. But in making such a choice, we give up the notion of a target population, so that the precision of our analysis for a particular group may decrease, given size and experimental budget. Or to state the problem in an alternative manner, for a given level of precision in estimation, the necessary budget for an experiment might increase substantially.

An alternative approach is to stratify on <u>exogenous</u> variables only and to approximate the goals of endogenous stratification by using predicted values of the endogenous variable.[8]

We shall consider the first example, sample truncation, since the issues can be seen quite clearly. Figure 1 represents the density of earnings with a truncation point T.[9] Suppose our aim is to sample people in the area of the distribution marked I. Now instead of using pre-experimental income with its associated problems, consider the use of "exogenous" income stratification, based on income predicted on the basis of exogenous variables, say from the regression equation

$$(13) \qquad \hat{Y}_i = Z_i\delta + \varepsilon_i$$

where the prediction is $\hat{Y}_i = Z_i\delta = Z_i\delta + Z(Z'Z)^{-1}Z'\varepsilon$. Note that ε_i still enters the last term through the product $Z_i'\varepsilon_i$. But for a sample of size N this term is of order 1/N, so that it quite rapidly disappears as the sample becomes large. The variables included in Z_i would be education, training, union membership, age, etc. We could then base truncation, so that problems which arise from the individual effect $\mu_i = \varepsilon_{it} - \eta_{it}$ being present in both periods no longer occur.

If the covariance between y_i and $\hat{y}_i$ were very high, we would have solved the problem. Then the predicted value would do almost as well as the actual endogenous variable. But for log earnings the R^2 of the regression is around .25 multiple correlation coefficients in the range of .25 to .60 are quite common for many cross section regressions in econometrics. Thus, if we use $\hat{y}_i < F$ as the truncation point we expect on average to do about 1.2 as well as pure random sampling in selecting $y_i < L$.

While this is an improvement, we might do even better by choosing a point $k < L$ as our sample truncation point. Perhaps a useful approach to the choice of k can be constructed as follows. Assume the benefit to estimation of the experimental effect has expected value of the form $V(y_i) = \beta/(y_i - \bar{y})^2$. That is, we expect to learn little about labor supply response from low income or high income individuals. On the other hand, cost is expected to grow linearly with income $c(y_i) = cy_i$. Suppose we then want to solve for the optimum truncation point k, given our knowledge that since we are using predicted income $\hat{y}_i$, the actual $y_i = \hat{y}_i + \varepsilon_i$ will differ. The optimization problem is

$$\text{(14)} \qquad \max_k \beta/(y_i - \bar{y})^2 \quad \text{s.t. } \Sigma c y_1 \leq C \quad \hat{y}_i = y_i - \varepsilon i \leq k \ .$$

We solve the corresponding expected value problem

$$\text{(15)} \qquad \max_k L = E(\beta/(\hat{y}_i + \varepsilon_i - \bar{y})^2) + \lambda_1 E(C - \Sigma c(\hat{y}_i + \varepsilon_i)) + \lambda_2(k - \hat{y}_i)$$

The form of the solution can be seen by assuming that the variable has been transformed to make the residuals approximately normal and that we center the data to set $\bar{y} = 0$. Then we choose k to

$$\text{(16)} \qquad \max_k L' = \beta/\left[\operatorname{var}(\hat{y}_i) + 1 - \frac{\left(\frac{k}{\sigma}\right)\phi\left(\frac{k}{\sigma}\right)}{\Phi\left(\frac{k}{\sigma}\right)} - \left(\frac{\phi\left(\frac{k}{\sigma}\right)}{\Phi\left(\frac{k}{\sigma}\right)}\right)^2\right] + \lambda(C - c\,\Sigma\hat{y}_i - \sigma\left(\frac{\phi\left(\frac{K}{\sigma}\right)}{\Phi\left(\frac{k}{\sigma}\right)}\right)$$

where σ is the standard deviation of the residual distribution. The first order conditions of equation (16) are straightforward and the problem can be solved straightforwardly on a computer since the constraint will be satisfied with equality and all the functions are monotonic in k. In this problem the gains over random sampling increase as the variance of the residuals decrease so that y_i and $\hat{y}_i$ are more highly correlated as we would expect. If the correlation becomes very small, we will be quite close to random sampling. But in many cases random sampling may be preferable to endogenous sampling, which as we have attempted to show, can lead to difficult problems in the analysis phase of an experiment.

IV. Self-Determination of Participation and Attrition

We have addressed in the previous sections a problem that we believe has been largely induced by experimental design and that we believe should be avoided. In this section we will address a major potential problem that we believe cannot in general be avoided but that can be corrected for without undue complication as long as it is not accompanied by induced endogenous stratification.

Suppose that it were possible to select a random sample of families from the population, or from a subset of the population (say with predicted income below some level). Of the families selected at random, some, when asked to participate in the experiment,will do so while others will elect not to participate. Even though a random sample is identified, those who choose to participate may not represent a random sample. In experiments to date there has been no systematic record kept of who when asked participates and who does not. Thus it has not been possible to identify systematic differences (and in particular unmeasured ones) between those who participate and those who do not, and, of course, if there were differences, there has been no way to correct for them. In the income maintenance experiments, for example, a procedure like the following was used. Each experiment was conducted within a single city or a small number of cities. All families within the city, or within some section of the city were canvassed to locate those with a few predetermined characteristics. In these experiments, income, race, age of family head, and number of dependents were attributes that determined eligibility. Those who were found to meet the eligibility criteria were asked to enroll in the experiment. Of those who did enroll, some were assigned to a treatment group, and others to a control group. It is the enrollment decision that concerns us here.

Suppose that instead of using a procedure like the above, we were to begin with an external source of data on families. The Census is a logical choice. Census data provide information on family income, race, whether the family is single or two-parent, education of family head, number of dependents, etc. Suppose that the known family attributes are represented by a vector of characteristics X. From families surveyed by the Census Bureau, a random sample could be chosen.

For simplicity, suppose the goal is to estimate a single treatment effect. Ideally we would like to randomly assign part of this randomly selected sample to a control group and others to the treatment group. Then after some time period, we would like to compare controls and experimentals, with Y the outcome of interest, using a simple analysis of variance model of the form

$$Y_i = \beta_0 + \beta_1 T_i + \varepsilon_i, \tag{17}$$

where T_i is an indicator variable with the value 1 for experimentals and zero for controls.

But suppose that not all of the random sample agrees to participate. Suppose that participation depends on X and a random disturbance term η in the following way:

$$(18) \qquad P_i = X_i\alpha + \eta_i \ ,$$

where P_i is an unobserved index variable with the property that individual i agrees to participate if $P_i > 0$. If Y_i and P_i are jointly normal with correlation coefficient ρ, and η is normalized to have variance 1, we know that the expected value of Y_i, given that individual i enrolls is given by

$$(19) \qquad E(Y_i|P_i > 0) = \beta_0 + \beta_1 T_i + \rho_{\varepsilon\eta}\sigma_\varepsilon \frac{\phi(X_i\alpha)}{\Phi[X_i\alpha]} \ .$$

Suppose that β_1 is estimated by least squares using the sample of participants and ignoring the last term in equation (3). Let the inverse Mills ratio $\phi(\cdot)/\Phi[\cdot]$ be represented by M_i. According to standard excluded variable arguments, if M is correlated with T, the least squares estimate of β_1 will be biased. As the sample of participants becomes large, the least squares estimate goes to

$$(20) \qquad \beta_1 + \rho_{MT}\rho_{\varepsilon\eta}\rho_\varepsilon \frac{\sigma_\varepsilon}{\sigma_T} \ ,$$

where ρ_{MT} is the correlation between M and T. If the treatment indicator T, however, is assigned randomly, then it will be uncorrelated with X and thus with M which is a function of X. Thus under these simple assumptions, the least squares estimate of the treatment effect will be consistent, as long as the assignment to control versus treatment groups is random. Each participant could be randomly assigned or each of those in the Census sample could be randomly assigned prior to enrollment, as long as at the time of enrollment, prospective participants did not know their assignment.

But the model as set out above hides by omission a potential major source of self selection bias. Suppose that if the treatment were given to all persons in the population, the responses would vary among them. It is clear that this is indeed the case (even after controlling for measured family characteristics). It seems plausible that the decision to participate will depend on the potential response. For example, it is often hypothesized that persons whose behavior is most likely to be affected will be most likely to participate, even though they do not know prior to enrollment whether they will be in the treatment or in the control group. This is the essence of the examples given in Section II-B.

The idea may be represented by a random effects model of the form

$$(21) \qquad Y_i = \beta_0 + (\beta_1 + b_i)T_i + \varepsilon_i = \beta_0 + \beta_1 T_i + b_i T_i + \varepsilon_i \text{ ,}$$

where from the perspective of the analyst b is random with mean zero. Using (21), the expected value of Y_i among participants is given by,

$$(22) \qquad E(Y_i | P_i > 0) = \beta_0 + \beta_1 T_i + (\rho_{b\eta}\sigma_b T_i + \rho_{\varepsilon\eta}\sigma_\varepsilon) \frac{\phi(\cdot)}{\Phi[\cdot]} \text{ .}$$

In this case, it is clear that the last term will be correlated with T_i, and a least squares estimate of β_1 would be biased.

Joint maximum likelihood estimation of (18) and (21), however, could be used to obtain a consistent estimate of β_1. The procedure is similar to the one proposed by Hausman and Wise [1979], except that the equations pertain to the response variable and participation, rather than to the response variable and attrition. In this case, there are two possible outcomes: Individual i doesn't participate with probability.

$$(23) \qquad 1 - \Phi[X_i\alpha], = P_{1i}$$

or individual i participates with response Y_i, with likelihood

$$(24) \qquad \Phi\left[\frac{X_i\alpha + \frac{\rho_{\eta b}\sigma_b T_i + \rho_{\varepsilon\eta}\sigma_\varepsilon}{\sigma_b^2 T_i^2 + \sigma_\varepsilon^2}\cdot(Y_i - \beta_0 - \beta_1 T)}{\left(1 - \left(\frac{\rho_{\eta b}\sigma_b T + \rho_{\varepsilon\eta}\sigma_\varepsilon}{\sqrt{\sigma_b^2 T_i^2 + \sigma_\varepsilon^2}}\right)^2\right)^{\frac{1}{2}}}\right]\cdot$$

$$\cdot \frac{1}{\left(\sigma_b^2 T_i^2 + \sigma_\varepsilon^2\right)^{\frac{1}{2}}} \cdot \phi\left(\frac{Y_i - \beta_0 - \beta_1 T}{\left(\sigma_b^2 T_i^2 + \sigma_\varepsilon^2\right)}\right) = P_{2i}$$

The likelihood function

$$(25) \qquad L = \sum_{i=1}^{N_1} \ln P_{1i} + \sum_{i=1}^{N_2} \ln P_{2i}$$

can easily be maximized to obtain estimates of β along with the other parameters of the model.

The other component of self-selection that seems unavoidable in social experiments is attrition. Some participants will inevitably drop out of the experiment before the treatment response is measured. To take advantage of individual specific characteristics that persist over time, it is also advantageous to observe participants for some period of time before the treatment becomes effective. This will lead to four equations of the form

$$\begin{aligned} P_i &= X_i\alpha + \varepsilon_{1i} , \\ Y_{1i} &= X_{1i}\delta + \varepsilon_{2i} , \\ Y_{2i} &= X_{2i}\delta + \beta_1 T + \varepsilon_{3i} , \\ A_i &= X_i\gamma + \varepsilon_{4i} , \end{aligned} \tag{26}$$

where Y_1 pertains to the response variable before the treatment period, Y_2 to the response variable during the experimental period and A is an unobserved indicator variable with the property that individual i leaves the experiment if $A_i < 0$. This system of equations can also be estimated readily with available maximum likelihood techniques. (See Venti and Wise [1981]).

REFERENCES

Goldberger, Arthur. 1980. Abnormal selection. Mimeographed. University of Wisconsin.

Hausman, Jerry A. 1980. The effects of time in economic experiments. Presented at the World Econometrics Society conference, Aix-en-Provence, August.

Hausman, Jerry A., and Trimble, J. 1981. Sample design consideration for the Vermont TOD use survey. Journal of Public Use Data 9.

Hausman, Jerry A., and Wise, David A. 1976. The evaluation of results from truncated samples: the New Jersey negative income tax experiment. Annals of Economic and Social Measurement 5: 421-45.

________________. 1977. Social experimentation, truncated distributions, and efficient estimation. Econometrica 45: 319-39.

________________. 1979. Attrition bias in experimental and panel data: the Gary income maintenance experiment. Econometrica 47: 455-73.

________________. 1980. Earnings effects, sample selection, and treatment assignment in the Seattle-Denver income maintenance experiment. Working Paper.

________________. 1981. Stratification on endogenous variables and estimation: the Gary income maintenance experiment. In Manski, Charles, and McFadden, Daniel, eds., Structural analysis of discrete data: with econometric applications. Cambridge: MIT Press.

Manski, Charles, and McFadden, Daniel. 1981. Alternative estimators and sample designs for discrete choice analysis. In Manski, Charles, and McFadden, Daniel, eds., *Structural analysis of discrete data: with econometric applications*. Cambridge: MIT Press.

Meier, Paul. 1978. The biggest public experiment ever: the trial of the Salk poliomyelitis vaccine. In Tanur, Judith M. et al., eds., *Statistics: a guide to the unknown*. 2nd edition, San Francisco: Holden-Day.

Venti, Steven F., and Wise, David A. 1981. Individual attributes and self-selection of higher education: college attendance versus college completion. Forthcoming in *Journal of Public Economics*. (Also under title "Test scores and self-selection. . . .", NBER Working Paper #709.)

______________ . 1982. Moving and housing expenditure: transaction costs and disequilibrium. NBER Working Paper #735.

FOOTNOTES

[1] We do not mean to disregard important problems which still remain. Questions of interactions may still arise, for example.

[2] For a further discussion of time effects in experimental design, see Hausman (1980).

[3] Of course, with only 2 periods this assumption is only a normalization.

[4] Manski-McFadden (1981) consider a similar question in attempting to minimize sample survey costs in a discrete choice model framework.

[5] The opportunity to do any type of nonparametric analysis is severely limited here because we do not have observations on the part of the sample that was truncated.

[6] If P_1 and P_2 are known, the analysis can be simplified somewhat. See Hausman-Wise (1981).

[7] The unknown aspect arises because there does not exist a straightforward model for assignment of E-level. Part of the assignment procedure involved qualitative judgments.

[8] This approach was used in the design of a survey for electricity use in Vermont by Hausman-Trimble (1981).

[9] We are assuming a common truncation point, although in the NIT experiment it depended on family size, which partly defines the poverty limit. But we can add varying truncation points to our analysis with no added complications.

Table 1. Reported Cases of Poliomyelitis[a]

Study Group	Study Population	All Reported Cases per 100,000
Placebo control areas		
Vaccinated	200,745	41
Placebo	201.229	81
Not inoculated	338,778	54
Observed control areas		
Vaccinated	221,998	34
Controls	725,173	61
Grade 2 not inoculated	123,605	53

[a] Numbers are from Table 1, p. 11 of Paul Meier (1978).

Figure 1. Selection Based on an Exogenous Variable

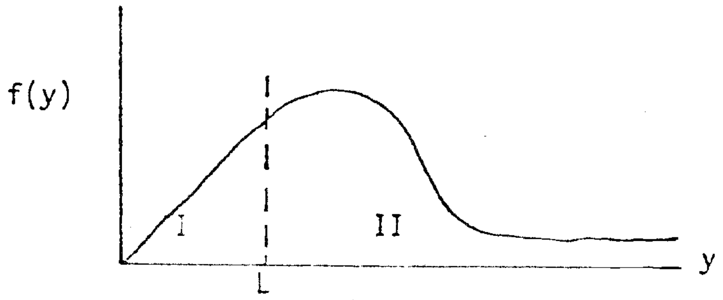

Figure 1