Evaluating the Methodology of Social Experiments

*Arnold Zellner and Peter E. Rossi**

In view of the many papers and books that have been written analyzing the methodology of the income maintenance experiments as well as other social experiments, it is indeed difficult to say anything entirely new. However, we shall emphasize what we consider to be important, basic points in the planning, execution and evaluation of social experiments that may prove useful in future experiments. The plan of the paper is as follows. In the next section, we put forward considerations relevant for evaluating social experiments. We then take up design issues within the context of static designs, while the following section is devoted to issues that arise in dynamic contexts, the usual setting for most social experiments. Suggestions for linking social experiments to already existing longitudinal data bases are presented and discussed. In both static and dynamic contexts, we discuss the roles of models, whether statistical or structural, and of randomization. Design for prediction of relevant experimental outcomes is emphasized and illustrated in terms of simplified versions of the negative income tax experiments. Finally, we present a summary and some concluding remarks.

Considerations Relevant for Evaluating the Methodology of Social Experiments

Since social experiments usually involve the expenditure of millions of dollars, resources that have alternative research uses and potentially great social value, it is critical that the experiments be conducted in a

*Professor of Economics and Statistics and Assistant Professor of Econometrics and Statistics, respectively, Graduate School of Business, University of Chicago. Michael A. Zellner provided helpful research assistance.

manner that is methodologically sound. The task of defining good or optimal methodology is difficult, however, because social experiments are multidimensional in nature, involving many disciplines and activities. Also, good experimentation involves creativity and innovation, which are difficult to define. We will discuss critical features that can be of vital importance for the success of social experiments.

A clear-cut statement of the objectives of an experiment is the first requirement of a good methodological approach. Poorly formulated objectives are an indication of poor methodology in general. If an experiment is purely for research, then the researchers have the responsibility for formulating its objectives. On the other hand, if the experiment has mainly policy objectives, then it is critical that researchers and relevant policymakers jointly formulate the objectives of the experiment.[1]

Once the objectives of an experiment have been clearly formulated, the second step involves a feasibility study, in order to determine how and if the objectives can be realized. This should include a review of previous studies and data, experimental and nonexperimental, relating to the objectives of the current experiment. It should also consider in detail the needed inputs for the proposed experiment. Usually subject matter specialists, well versed in the subject to be investigated,[2] survey experts, and design statisticians will be required. Most important is the development of an operational approach that is capable of realizing the objectives of the experiment.

If the objectives involve the production of results in a short time, the feasibility study may indicate that calculations using nonexperimental data are all that can be done. On the other hand, if a social experiment seems feasible, its design and costs should be explicitly developed. Finally, the quality of both the research team and the managerial or administrative personnel is of key importance.

In the feasibility study, it is desirable that calculations be performed to provide preliminary estimates of important effects.[3] These rough calculations provide important order-of-magnitude estimates that can be quite useful as background information in evaluating experimental designs. Last, it is usually good practice to execute a "pilot" or "test" trial of the experiment, just as survey questionnaires are subject to pretests. Such pilot experiments can reveal many unexpected results and aid in the redesign of the "final" experiment.

The quality of measurements is a third key issue in the evaluation of the methodology of social experiments.[4] If measurements are of low quality, the results of an experiment are of dubious value. Are all appropriate and relevant variables being measured? Are the measurements afflicted by response and recall biases? Do subjects misrepresent data for various reasons? Are Hawthorne effects present? Checks on the validity of the basic data provided by an experiment must

be pursued vigorously in a good methodological approach to social experimentation. This requires that data specialists and those familiar with measurement methodology be involved in the execution of a social experiment.

Fourth, as stated above, outstanding subject matter specialists are required, in order to ensure that the methodology of an experiment is appropriate.[5] An experiment usually involves subjecting experimental units to important changed conditions. Since their responses to the changed conditions are usually adaptive and dynamic in nature, care must be taken in choosing a model that can represent such responses and serve as a basis for choice of an experimental design. Designs can be chosen not only to estimate effects from given models but also to provide data that are useful for testing uncertain features of a model brought to light by experts' analyses of existing theory and models. For example, such considerations may involve use of a model with several equations rather than a single equation for, say, labor supply. If the multiple equation model is appropriate, a design based on a single equation is inappropriate and can lead to erroneous conclusions.[6]

Fifth, the design of the experiments and other statistical issues are basic to a good methodological approach. If the objectives of an experiment involve generalization of the results to an entire population, then the sample of experimental units has to be a sample from the relevant population.[7] The relevant population must be carefully defined with respect to spatial, temporal[8] and other characteristics. Further complications arise from the inability of experimenters to require participation in an experiment. Those volunteering to participate may possibly be different from those not willing to participate and if so, the experiment may be subject to selection biases. (See Duan et al. (1984) for an evaluation of models that attempt to correct for selection bias.) Further, there is the problem of attrition.[9] It is important to do everything possible to keep the attrition rate low. Sampling the dropouts and using the results of analyses relating to these samples is one way of checking on the importance of attrition bias and of correcting for it. Constant vigilance with respect to possible sources of bias and the use of every means possible to avoid such biases are characteristic of a good methodological approach.

Assigning units to treatment and control groups at random is considered good practice by most experimenters. However, good randomization procedures depend on an intimate knowledge of the model generating the observations, as Jeffreys (1967, p. 239ff.) and Conlisk (1985) have demonstrated. For example, Conlisk (1985) has shown that effective randomization when treatment effects are additive is not effective when treatment effects enter a model nonlinearly. Thus, how one randomizes depends on what one knows about features of a model for the observations—for example, see Rubin (1974). Also, a randomized

design for a static model may not be effective if the static model misrepresents dynamic responses to "treatments" and other variables. Last, it is worthwhile to emphasize not only precision in design but also balance and robustness of design, as Morris (1979) has emphasized.

Sixth, a successful social experiment requires good managers and administrative methods. A good research manager will be invaluable in scheduling operations of a large-scale social experiment, keeping control of costs, instituting good data management procedures, and, most importantly, guiding the project so as to raise the probability of its success in meeting its objectives. Good management involves not only selection of appropriate researchers and other personnel but also surveillance of the project to ensure that researchers are pursuing the stated objectives of the experiment. There usually is a great danger that researchers may get involved in tangential problems and issues and possibly provide the right answer to the wrong problem, a statistical error of the third kind.[10]

As the experience with the negative income tax experiments has indicated, data collection and data processing costs have been a large fraction of total costs of social experiments.[11] Thus, it is important to put a great deal of emphasis on the design of efficient computerized data management systems and on ways to record basic data directly into computerized data bases.

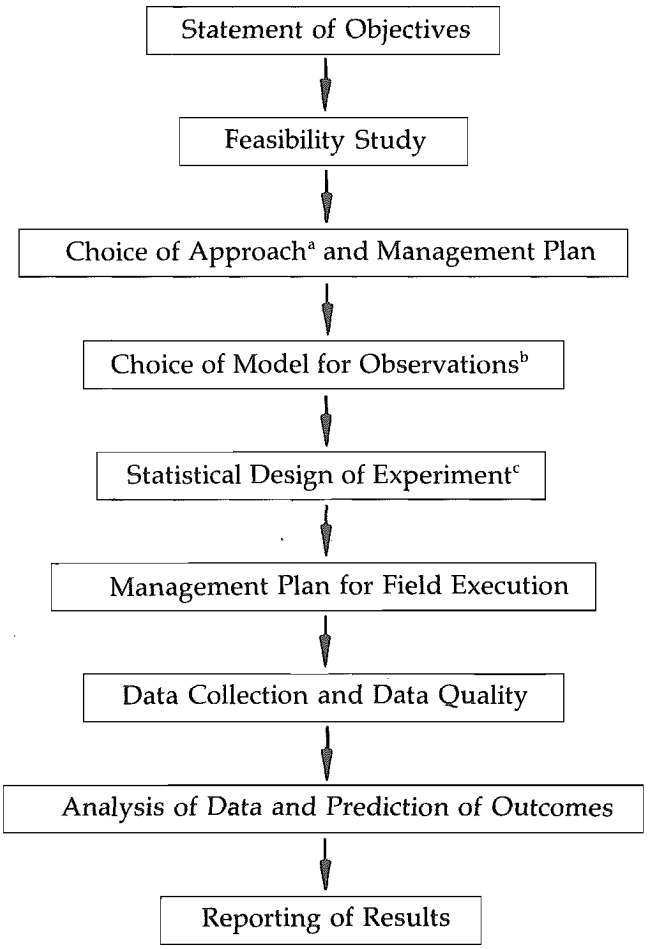
Seventh, good experimental research methodology involves concentration on the prediction of observable outcomes of an experiment and on establishing reproducible results. Predicting the observable outcomes of a social experiment, for example the costs of variants of a negative income tax program, might be the main objective of a social experiment. If so, the experiment should be designed to provide accurate enough predictions of the costs of these variants along with measures of predictive precision. Calculations yielding these predictions should be reproducible. However, the final proof will be the extent to which the experiment's predictions agree with the actual cost of a negative income tax program, if instituted. That is, good experimental methodology results in predictions that are verified in practice.

Finally, the results of an experiment must be well reported. Careless reporting of experimental results can lead to the adoption of incorrect policies, particularly when the experiment's objectives are not one-to-one with policymakers' objectives.[12] Good methodology requires great care in reporting the results of experiments with attention paid to their uncertainties and measures of precision. Results must be reported in a probabilistic framework that policymakers can understand.

The schematic diagram may be useful in providing an overall appreciation of the process of social experimentation, although it should be recognized that possibly important feedback loops have not been in-

cluded. For example, at various stages in the execution of an experiment research findings may indicate a need for changes in objectives, the model chosen for the observations, or other aspects of the experiment. While some of these eventualities may be taken into account in sequential designs, taking appropriate measures to deal with the unforeseen is an important element of a good methodological approach. (Watts and Bawden (1978) provide an account of some surprises that actually occurred in the experiments.)

Elements Involved in Planning and Executing Social Experiments



[a]It is assumed that a social experiment approach is feasible and selected.

[b]It is assumed that a relevant population has been defined.

[c]The statistical design can involve checks on the adequacy of alternative models for the observations and reflect what is learned from "pilot" or "test" trials of the experiment.

Design and Analysis in a Static Framework

A substantial literature has developed on the econometric problems associated with the design of social experiments and the analysis of experimental data. We shall not attempt an exhaustive summary or critique of the econometric literature on social experiments.[13] Rather, in this section, we shall discuss some key features of design of social experiments in the context of a simple "static" setting. The social experiment will take place during one experimental "period," which is assumed to be long enough for households to respond fully to the experimental treatment. The simple static setting will allow us to stress statistical problems common to most social experiments without detailed exposition of complicated structural models. Of course, use of a static model abstracts from some important dynamic considerations in both the design and analysis of experiments, which we will take up in a later section.

Review and Critique of Existing Design Procedures

The design of social experiments involves three distinct sorts of decisions: 1) the choice of experimental population, 2) the choice of the design space or range of possible treatments, and 3) the allocation of subjects (typically, households) to various treatment and control groups. In all major social experiments, specific sites (most often SMSAs) have been chosen for an experiment. On each site, the eligible population is determined and a sample from this population is invited to participate in the experiment. High administrative and field interview costs are cited as reasons for using a site-based approach.[14] The goal of a representative site (Middletown, USA) has proved elusive. As a result, the experimental population is frequently not representative of the target population for a national program. In order to extrapolate the findings of the experiment to a national scale, some sort of structural model must be assumed.[15]

National probability samples should be considered for future experiments. The problems with administration and field operations of a national experiment appear to have been exaggerated by the planners of the negative income tax experiments. A number of national survey organizations have routinely conducted national surveys since the 1960s. (The National Opinion Research Center at the University of Chicago and the Institute for Social Research at the University of Michigan are examples.) Differences in local welfare laws make it all the more necessary to diversify the sample to include more than a handful of sites.[16] In the next section, we propose the formation of an ongoing panel based on a national sample frame for use in social experiments. It

should be noted that marketing researchers have used both the national sample approach (Nielsen surveys) and the site-based approach (most often used in the test-marketing of products) to evaluate the effects of changes in advertising and the response of consumers to new products.

Once a target population is selected, the designers of the experiment must select the range of possible treatments. In the negative income tax experiments, the possible combinations of tax rates and support levels determine the design space. The possible range of treatment variables appears to have been selected in an ad hoc fashion through a combination of political compromise and personal judgment. Describing the New Jersey experiment, Conlisk and Watts write:

> The problem, then, was one of specifying a sample in the three dimensional design space of (g,t,w) triplets [support, tax rate, and wage rate]. Sampling was restricted to a region within the design space which provided substantial variation . . ., but kept to (g,t,w) combinations of actual policy interest . . . So the design problem reduced to finding optimal numbers of families . . . to allocate to each design point.[17]

Keeley and Robins (1980b, p. 328) point out in their excellent critique of the design of the Seattle-Denver income maintenance experiments that it is not clear how to "specify the design space" and, more importantly, that "efficiency may be increased by changing the design space for a given response function." The problem of choice of design space is not unique to either the Conlisk/Watts response surface approach (used in all negative income tax experiments) or the ANOVA model approach (used in the design of the Housing Allowance Demand Experiment) to experimental design. Neither approach gives adequate guidance in the choice of the range of treatments.

The designs used in the negative income tax experiments tended to be overly conservative with less variation in treatment than is desirable, particularly when the goal of the experiment is to estimate quantities that are imprecisely known to start with. For example, in the Seattle-Denver experiment, marginal tax rates of between 0.5 and 0.8 were used. The range of these rates is very limited and does not include either the high or the low tax rates that might be expected to produce the most or least labor supply reduction. It is precisely from such extreme experimental conditions that most can be learned about model parameters and model adequacy. In the Seattle-Denver experiments, a tax rate of 0.3 and a support level of $3800 would have had a grant break-even level of $12,667, only slightly higher than the highest break-even level in the study ($12,000).

The distribution of treatment points over the feasible set of treatments is critical for ensuring appropriate precision in the estimation of

treatment effects and for testing the model specification. Without an adequate range of treatment levels, precise treatment effect estimates can only be obtained from huge samples, beyond the means of social experimenters. Perhaps most importantly, without a fairly uniform distribution of treatments across the design space it is difficult to test for model misspecification by comparing alternative models. We find little attention directed toward this important problem in the social experimentation literature.

Allocation of participating families across treatment groups has received considerable attention. With few exceptions, economists have rejected a classical analysis of variance approach in favor of a model-based approach developed by Conlisk and Watts. (See Conlisk and Watts (1979) for a full description of this technique.) A response surface-based allocation of households to treatment groups was used, usually with modifications, in all the negative income tax experiments, the health insurance experiment, and several time-of-day electricity pricing experiments. The heart of the response-surface approach to design is a demand equation: in the case of the negative income tax experiment, a demand for leisure equation is used. Given a specification of the demand equation, optimal allocations of households across treatment groups are derived. The goal of the experimental design procedure is assumed to be the maximization of the precision of estimation of the coefficients of the response model, subject to a budget constraint. Optimal allocations from the demand equation model provide non-orthogonal experimental designs, which should provide greater precision in estimation of key model parameters than traditional orthogonal designs as well as samples with lower experimental cost.

A critical assumption in the application of the response surface models is that experimental observations may be more costly than control observations both in terms of benefits and administrative costs. In their excellent critique of the Conlisk/Watts model, Keeley et al. (1980, p. 328) indicate that in the Seattle-Denver experiments experimental observations were assumed to be four times as expensive as controls, when the actual cost ratio was 2.3. Because of the asymmetry of observation costs, the response surface approach yields non-orthogonal and non-randomized experimental designs. Keeley and Robins (1980a, b) and Hausman and Wise (1985) have emphasized the severe problems produced by the endogenous stratification induced by this cost function. It is difficult to understand this preoccupation with cost-effective designs when most of the significant costs of negative income tax experiments are fixed or at least constant across households. Typically, the data processing, field operations and analysis budgets for negative income tax experiments far exceeded the total benefits paid out. (See, for example, the cost data in footnote 11.)

Perhaps the most serious defect of the response model approach is the extreme sensitivity of the optimal designs to model misspecification. As Conlisk (1973) and Aigner (1979a) have reported, optimal designs for one response function can be very suboptimal for other response functions. Given the considerable uncertainty about the appropriate specification of labor supply behavior, a good design procedure should incorporate some robustness to departures from model assumptions. Model misspecification can include incorrect functional form for the demand equation, measurement errors in the independent variables, omitted or latent variables, incorrect distributional assumptions (outliers, sample truncation and censoring), sample selection bias, and incorrect dynamic specifications. Due to the inherent difficulties in measuring income and wage rates and the field operations mistakes made in the negative income tax experiments, the problems of accounting for measurement error biases are particularly important. Measurement error problems are further compounded when the design is stratified based on an endogenous variable that is measured with error. Explicit consideration of measurement errors is a necessary condition for optimal design in these situations.

A natural way to produce robust statistical designs is to use randomization procedures, by which households are randomly assigned to treatment groups according to simple or stratified random sampling procedures. Randomization procedures have been widely used in experimental design in medical, psychological and educational research for many decades. Based on both practical experience and extensive theoretical research (Fisher, 1925 and Kempthorne, 1952) randomization procedures have been shown to have great value in reducing bias in the determination of experimental effects when response models are misspecified. As Hausman and Wise (1985) and Morris (1979) have pointed out, to the extent that unobserved variables are correlated with observed variables over which the design is randomized, the effects of model misspecification are mitigated.[18]

The severe problems that have plagued the response model approach to design have prompted some economists to advocate designs based on the simplest sort of analysis of variance (ANOVA) model. (See, for example, Hausman and Wise, 1985.) Optimal design in an ANOVA framework requires an orthogonal layout of treatments and a random assignment of participants to the various treatment groups (each group corresponds to a row of the X matrix in the Conlisk/Watts model). Of course, the ANOVA model is a special case of the general linear response model used by Conlisk and Watts. It appears that the chief benefits of the ANOVA approach are simplicity of analysis and randomization over participant characteristics, which avoids the problem of endogenous stratification. However, the ANOVA approach is sensitive

to model misspecification, as pointed out by Conlisk (1985), requires many observations when there are large numbers of cells, and, most importantly, cannot be used to generalize in important ways from the experimental experience. That is, ANOVA models are designed to test for experimental effects and cannot be easily adapted for predictive purposes. Also, given the problems of a site-based sample and participation and attrition biases, it is highly unlikely that experimental data will be representative of the national target population for a social program. Thus, a response model would have to be built to extrapolate the ANOVA experimental findings to a national scale.

An essential problem with both response function and ANOVA approaches to experimental design is that research objectives of a study are not explicitly included in the objective function used to determine the optimal design. If the objective of a negative income tax experiment is to estimate accurately the cost of a national program, the objective function should be formulated to measure the precision of cost estimates. Similarly, if the goal is to refine estimates of substitution effects, the precision of these estimates should define the objective function.[19] The usefulness of the current experimental design techniques can be gauged by noting that we know of no social experiment in which the original design model was used in the analysis of experimental data. In all of the negative income tax experiments the original response model was discarded in favor of more restrictive labor supply functions, which have drastically fewer parameters.

The unusual role of controls in social experiments also involves difficulties with current design techniques. A control subject is defined in the classical experimentation literature as a subject who received no treatment. In the negative income tax experiments, many control families received current AFDC benefits (of course, the investigators could not ask these households to give up income support altogether) while other families received no welfare benefits of any kind. These control AFDC families are receiving a different treatment, not the null treatment. In fact, the control households are treated in most analyses simply as additional experimental households with different tax rates and disposable income. Control households are lumped into the sample to "increase estimation efficiency."[20] Investigators often perform some sort of pooling test to see if controls and experimentals can be lumped together. The question of whether controls are necessary in any fundamental sense except as low-cost observations has not been adequately addressed. One could also ask whether experimental observations are necessary and whether the existing experimental variations in prices and income are sufficient to estimate household response functions precisely.

A design procedure touted to be optimal must properly specify the

objective function of the experiment as well as determine if the experiment is necessary. In the next section, we outline a decision-theoretic approach that provides a workable solution to this problem.

A Decision-Theoretic Approach

The key to developing a useful experimental design is a well-defined and meaningful objective function. Clearly defined objectives are critical in planning for a useful experiment. By forcing both the contracting agency and the investigator to specify clear and quantifiable objectives, it is possible to determine accurately whether an experiment is necessary and to produce a design that is able to discern treatment effects. Two main objectives were pursued in the negative income tax experiments: 1) computation of the net program cost of a national negative income tax program[21] and 2) estimation of the national labor supply response (work disincentive) to proposed negative income tax programs. The estimation of the labor supply response is a less ambitious goal than the costing of a national program and comes closest to the goal of most of the principal investigators.

It is also crucial that the results of a study be formulated in a way that can be effectively communicated to policymakers. We have found the research memoranda and papers on the negative income tax experiments to be very difficult to decipher even for readers with considerable econometric expertise. The results of statistical analyses are often reported without standard errors or evidence of diagnostic checking of any kind, without the number of observations in the estimation sample, and with poor labeling of tables and diagrams. More importantly, however, the extreme emphasis on point estimation and significance tests results often leads to misleading reports. For the policymaker interested in the costs of a national negative income tax, it is not sufficient to supply a point estimate of total costs. Some measure of the uncertainty in that point estimate due to estimation and possible specification error must also be supplied. It is extremely difficult to convey the uncertainty in point estimates just by supplying the estimate and the standard error of estimate. It is more useful to supply an interval and some probability statement about that interval. For a policymaker evaluating a negative income tax program, a statement such as "Given the information obtained in this experiment, we can say with a probability of .9 that the net program cost falls between A and B," is useful and understandable. Such a statement is one aspect of the predictive density of program costs, given the information available at the time of the report. The predictive probability density function expresses information about future costs on the basis of past sample and prior information with a due allowance for parameter uncertainty.

The policymaker can be supplied with various summary measures of the predictive density of costs, including the probabilities associated with various prediction intervals and measures of dispersion (variance, standard deviation and interquartile range) as well as plots of the density. In this way the policymaker can be assured that the accuracy of information available on key response parameters has been taken into account. Uncertainty with respect to the form of the response model can also be quantified, a problem treated at the end of this section.

The failure to report estimates of the precision of national cost estimates in the existing literature is all the more disturbing given that the national labor supply response may be very imprecisely estimated. To illustrate this point, let us examine the Keeley et al. (1978b) estimates of total labor supply response to various negative income tax programs. Keeley et al. estimate the hours response of heads of households and wives by applying a fitted labor response equation to a national probability sample of households derived from the Current Population Survey. By adding up the estimated hours supplied for each record in the file, Keeley et al. are able to produce national estimates of the work incentive/disincentive effect of various negative income tax programs. The labor supply response is figured in with other welfare and tax effects to impute the net cost of negative income tax programs. Their analysis depends critically on the quality of their estimated labor supply response functions.[22] Careful examination of their fitted response functions reveals that the fits are very poor and the coefficients measuring income and substitution effects are imprecisely estimated.[23] For example, in the equation for husbands, the standard error of regression is 720 hours per year. The mean number of hours worked per year before the negative income tax program is 1,999 for husbands. This suggests that the error variance in the labor supply response relationship is very large. (It also suggests that the model may be misspecified.) Even without taking account of estimation error, we would expect that predictions of supply responses to a national program would be very imprecise. In addition, the coefficients of the key variables are very imprecisely estimated. In the equation for husbands, not a single coefficient is estimated to within one significant digit of precision!

The results of microsimulation presented in table 7 of the Keeley et al. (1978b) paper do not include measures of precision. This gives the reader a false sense of the accuracy of these results. To illustrate the potentially enormous standard errors of prediction for these numbers, we will undertake some approximate standard error calculations. To obtain these crude figures, we must make many simplifying assumptions because we do not have access to the experimental and national data. However, the assumptions that are made bias downward our estimates of the standard error of prediction. Keeley et al. report an

average reduction of 19 hours per year for husbands in a negative income tax program with support equal to 75 percent of the poverty level and a 50 percent tax rate. To put this figure in a form useful for computing program costs, we express all estimates as national totals. An average reduction of 19 hours implies a total reduction of 756 million hours per year for the total labor force of husbands. Keeley et al. have computed the total hours figure by summing up individual estimates as follows.

$$\hat{H}_{tot} = \sum_{i=1}^{N} \hat{H}_i \tag{1}$$

To derive the variance of the prediction error, $H_{tot} - \hat{H}_{tot}$, we must make some assumptions about the prediction errors for each individual's equation. We assume that the parameters of the response function are known and concentrate on the source of variability from the inherent randomness in the labor supply relationship, that is its error term. The calculation of the prediction error variance depends critically on the assumption used for the joint distribution of the labor supply of husbands. If it is assumed that the error terms of each husband are independent, the results are radically different from those obtained in the case in which even a small amount of dependence is allowed between units. To simplify the calculations, we assume that the N × N error covariance matrix has a simple patterned structure,

$$\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho & \cdot & \cdot & \cdot & \cdot & \rho \\ \rho & 1 & \rho & \cdot & \cdot & \cdot & \rho \\ \cdot & \rho & \cdot & \cdot & & & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & & & \cdot & 1 & \rho \\ \rho & \rho & \cdot & \cdot & \cdot & \rho & 1 \end{bmatrix} = \sigma^2[(1-\rho)I_N + \rho \underset{\sim}{1}\underset{\sim}{1}'],$$

where $\underset{\sim}{1}$ is an N × 1 vector with unit elements. Under these simplifying assumptions, the variance of the aggregate prediction error is given by,

$$\text{Var}\,(H_{tot} - \hat{H}_{tot}) = \sum_{i=1}^{N} \text{Var}\,(H_i - \hat{H}_i) + \sum_{i \neq j} \text{cov}\,(H_i - \hat{H}_i, H_j - \hat{H}_j)$$

$$= N\sigma^2 + N(N-1)\,\rho\sigma^2 = N\sigma^2\,(1 - \rho + N\rho).$$

The standard error, $\sigma = 720$ of the estimated tobit equation in table 3 of the Keeley et al. paper and an assumption about the value of ρ can be combined to produce a prediction error variance estimate. With $\rho = 0$, $N = 39.8$ million, and $\sigma = 720$, the prediction standard error is 4.54

million hours, which would produce a tight prediction interval around the point estimate of −756 million hours. If we allow for even a small amount of dependence by assuming $\rho = .01$, the prediction standard error increases to 2.86 billion hours. It should be emphasized that we are neglecting estimation error in these calculations by assuming that the conditional mean function is known to the analyst. The sensitivity of these calculations to what is assumed about the value of ρ is striking.

The fundamental goals of social experiments have been predictive. Either Bayesian or classical prediction techniques can be used to produce predictions and measures of precision. These same techniques can be used to determine the optimal design of experiments. Before the experiment is undertaken, some prior information is available on the key response parameters. In the case of labor supply, a number of studies with nonexperimental data[24] have produced both substitution and income elasticity estimates. These estimates can be combined to form a prior distribution for the response parameter vector, $p(\theta)$. The predictive density of hours can be computed using this prior[25] and compared to the predictive density which would be obtained after the experiment. The before and after predictive densities can be compared to determine if a given experiment has sufficient information value to warrant undertaking it. Peck and Richels (1986) use a similar decision-theoretic approach to indicate how to decide upon future research on the acid rain problem. Stafford (1985) also proposed a decision-theoretic approach for determining if negative income tax experiments are useful. Stafford proposed a social utility function and suggested that the information value of social experiments be measured via social utility. We avoid the problems associated with postulating a social welfare function and focus on the narrower goal of evaluating predictive accuracy.

Comparison of predictive densities can be accomplished by computing various scalar measures of differences in the distributions. Interval and probability estimates may be the most useful computations. For example, it may be that a 90 percent probability interval for the national costs of a negative income tax might include a wide range of values including cost greatly above and below current welfare costs. It may be the case that a given experimental design may sharpen up this interval to the point that policymakers may feel comfortable with the point cost estimate. One suspects that these types of calculations, when applied to the designs used in previous negative income tax studies, would suggest that the experiments had little informational value. Other summary measures that may be considered are variances and other moments and the divergence of two densities.

We have emphasized in earlier sections that uncertainty about model specification is one of the most serious problems confronting the design analyst. The frequent assumptions of log-normality and linear

functional form of the labor supply function can easily be challenged. Problems with sample selection bias from the participation decision, attrition, and missing values also plague social experiments.[26] For example, Ferber and Hirsch (1979) point out that only 345 data points out of the more than 1300 enrolled households in the New Jersey experiment were actually used in estimating labor supply response. In the Seattle-Denver income maintenance experiments, approximately 1600 out of the total 2600 husband-wife households were used in fitting the labor supply equation. Useful optimal design procedures must consider the problem of optimal model selection and discrimination between alternative parametric models — see Box and Hill (1967), Covey - Crump and Silvey (1970), Guttman (1971) and O'Hagan (1978).

Given the considerable uncertainty regarding model specification, we are puzzled about the lack of discussion of predictive validation of models in the research reports of social experiments. The experimental data often contain numerous subsamples corresponding to different sites or different time periods or different treatment groups which could be used for validation purposes. For example, labor supply response functions fitted to Denver data could have been used to predict responses for the Seattle sample. If the labor supply function is well specified, the error terms should only contain random shifts due to tastes and other omitted characteristics of the households, and the prediction errors should follow the assumed error distribution in the model specification. If the response function cannot reliably extrapolate the results from one site to another, it is unreasonable to expect the same specification to be useful in predicting response to a national program.

A useful and easily generalized approach to model selection involves calculation of posterior probabilities of models in a Bayesian framework. Consider two different probability models for labor supply response, H, $p_1(H|\theta_1)$ and $p_2(H|\theta_2)$ where $\theta_1$and θ_2 are parameter vectors. For a given data set, we can compute the posterior probability of each model as follows:[27]

$$\Pr(\text{model i}|\text{data}) = \int l_i(\theta_i|\text{data})p_i(\theta_i)\Pr(\text{model i})d\theta_i/G \qquad (2)$$

where G = Pr(model 1|data) + Pr(model 2|data). The key ingredients in (2) are the likelihood function for each model, $\ell(.)$, the prior density for model parameters, $p_i(.)$, and the prior probability of the model. For comparison of models, we note that this approach does not require that the models be nested or that the models exhaust the set of plausible models. We do not adopt an accept/reject philosophy which eliminates models from future consideration even if the information in the data is insufficient to distinguish between the models. For many problems of practical interest, posterior model probabilities can be calculated without resort to asymptotic approximations.

One of the principal econometric problems encountered in the modeling of labor supply stems from the mass point at zero hours supplied in the empirical hours distribution. The analysts in the Seattle-Denver experiment used the truncated normal regression or tobit model to account for the massing of hours at zero:

$$\text{TOBIT model: } H_i^* = \underset{\sim}{X}_i' \underset{\sim}{\beta} + \epsilon_i \qquad \epsilon_i \sim \text{iid } N(0,\sigma^2)$$

$$H_i = \max (H_i^*, 0).$$

The tobit model can be considered as a special case of a two-equation model in which the first equation predicts labor force participation and the second equation gives labor supply response conditional on labor force participation.[28] A simple "two-part" model can be constructed by writing the density of H as consisting of a discrete and continuous part.[29]

$$p(H) = PR(H > 0)p(H|H > 0). \tag{3}$$

The participation equation can be a simple logit or probit binomial response model and the conditional distribution of hours for those in the labor force can be modeled with a simple regression function as follows:

$$\Pr(H_i > 0) = F(\underset{\sim}{x}_i'\underset{\sim}{\beta}) \tag{4}$$

$$\underset{\sim}{H} = Z\underset{\sim}{\theta} + \underset{\sim}{\epsilon} \quad \text{for } H_i > 0. \tag{5}$$

The distributional assumptions and parameter restrictions behind the tobit model are difficult to verify directly with data. One possible approach would be to fit the bivariate normal sample selection model of Heckman (1979) which nests the tobit specification. A significance test could then be performed on the parameter restrictions. However, the nested model hypothesis-testing approach often results in rejection of restricted models in favor of an unrestricted "super" model. Unfortunately, there is no clear connection between the tests of restrictions and the predictive performance of the models. It may well be that a tobit restrictions test may yield a rejection even though predictions from the tobit model may not differ much from a more complicated model. In fact, due to estimation error the simpler models may have a smaller out-of-sample mean squared error of prediction. Thus, predictive comparisons of alternative models should form the basis of model selection.

The normality assumption at the heart of the tobit and Heckman models is more difficult to check. The "two-part" model in equations (4) and (5) is more flexible and does not require strict normality assumptions. It would be of great interest to compare the tobit and two-part specifications. Equation (2) can be used to compute posterior model probabilities in much the same spirit as Rossi (1985). The integration

over the model parameters necessary to compute the posterior probabilities can be performed with Monte Carlo numerical integration techniques as in Rossi (1985), or asymptotic normal expansions of the posterior distribution of the parameter vector can be used.[30] It should be noted that the posterior model probabilities can be used to average predictions from alternative models—see, for example, Geisel (1975, p. 229). We can hedge our bets on model specification by carrying along a small set of competing models until (if ever) one model specification dominates.

By stressing the predictive objectives of social experiments, it is possible to solve both optimal design and estimation problems in a manner consistent with a study's objectives. As a general rule, past analyses of social experiments have not stressed predictive validation or useful reporting of national cost and response estimates. We hope our suggestions will motivate a rethinking of the statistical methodology for use with experimental data.

Design and Analysis in a Dynamic Framework

Dynamic aspects of economic behavior have received increasing attention in both theoretical and empirical literature. Labor economists were among the first to stress the importance of dynamic economic models. Recent empirical labor economics often uses time series or panel data to explore dynamic econometric models. Much of this effort to understand economic dynamics was exerted during the time of the major social experiments. As Griliches[31] has pointed out, "theory . . . could be changing exogenously, thereby making the experiment less interesting than originally." Future experiments designed to gauge the response to social programs will have to be designed to illuminate some of the dynamic aspects of economic behavior.

In labor economics, dynamic models of labor supply have been developed to explain the life cycle pattern of labor/leisure allocations as well as the patterns in the spells of employment and unemployment. As many have pointed out, the labor supply schedule derived from a life cycle model of labor supply can differ substantially from the model based on a one-period demand-for-leisure analysis. Changes in income support programs can also affect the search for new jobs and the durations of periods of employment and unemployment. Much of the labor force is under implicit or explicit contracts which would be affected by changes in transfer programs. The stochastic process governing labor force participation is much more complicated than the simple Bernouli trial models behind common statistical specifications. The assumption that workers form rational expectations suggests that macro-level

variables, which are useful in predicting the future course of output and wages, should be included in micro labor supply functions.

On a practical level, one of the most important dynamic aspects of social experiments is the problem of experiment duration. Most social experiments' test programs are intended to be implemented on a "permanent" basis or certainly for a much greater duration than the experiment. It seems obvious that households will react differently to permanent rather than transitory changes in wages and income. Moreover, experiments may be influenced by business cycle effects. The negative income tax experiments are of a short duration, typically lasting fewer than five years.[32] As an extreme case, the Housing Allowance Demand Experiment featured treatment cells in which households more often than not had to find new housing meeting rather arbitrary quality standards in order to qualify for small rent subsidies over only a two-year period. Other social experiments, notably some of the medical treatment experiments, have been conducted over much longer periods. In experimental situations in which treatment is limited to short durations, the challenge for the analyst is to calculate unobserved long-term effects.

As in our discussion of static models, we shall focus on the labor supply response relationship in the discussion of dynamic models. In the simple static model only current wages and income enter into the response function. Obviously, dynamic models enlarge this specification to include lagged response and input variables.[33] However, given the considerable debate over the appropriate dynamic theory of behavior, it may be wise to design in the context of an unrestricted transfer function model:

$$H_t = v_1(B)w_t + v_2(B)y_t + a_t$$

where w_t is the wage rate, y_t the income series, a_t follows a linear time series model, B is the backshift operator and $v_1(B)$ and $v_2(B)$ are rational functions of B. Of course, our ability to estimate lag structures can be severely limited by a short duration of an experiment. In such a situation, it is not clear that use of a static model will yield reliable results.

Time series analysts have developed a host of techniques for dealing with the adaptation of social and physical systems to environmental changes. In the intervention analysis pioneered by Box and Tiao (1975), time series models that allow for a wide variety of adjustment behavior are developed. In financial economics, critical events such as mergers or changes in government regulation are routinely studied with time series regressions and residual diagnostics in so-called "event" studies.

In order to design effective experiments for understanding complicated dynamic phenomena, we believe that a longitudinal design philosophy may be fruitful. In the labor supply problem, we observe

very substantial individual variation coupled with complicated dynamic and program duration effects. The "one-shot" experiment may not be the most effective use of experimental resources. We propose a longitudinal design scheme in which an ongoing panel of households is used as the population for a large number of smaller-scale experiments. Many economists would agree that the lack of well-collected panel data on the household level is a critical problem for economic research. The few panel data studies available[34] have spurred tremendous interest and basic research in labor economics. It is our view that social investment in panel data collection on a permanent, ongoing basis has an extremely high social rate of return.

The existence of a reasonably large panel of households for which detailed information on most key aspects of household behavior is available radically reduces the start-up and overhead costs of experiments. It would not be necessary to perform huge screening operations to identify eligible experimental populations. A long time series of pre-experimental data would be available for each household so that investigators would not have to rely on recall. A national sample frame would be ensured at low cost. Numerous checks and mini-experiments could be performed to reduce measurement errors or, at least, to understand the properties of measurement error. It would not be necessary to train and organize a field interviewing staff from scratch as was done in the negative income tax experiments. The longitudinal and ongoing nature of such a project will also force those implementing the study to design carefully for the coordination between field operations and analytical database management, an area in which many unanticipated difficulties were experienced in the negative income tax experiments.

Many analysts have noted the tremendous individual variation in labor supply response. The poor fit of cross-sectional labor supply functions is usually attributed to large, unmeasurable individual taste effects. In a longitudinal design scheme, households could serve as controls for themselves by alternating experimental and control treatments. The diagram below indicates a possible design layout for longitudinal design.

	Period			
Household	1	2	3	4
1	O	O	O	O
2	X	O	X	X
3	O	X	X	O
4	X	X	X	X

X denotes treatment and O denotes no treatment. In this design, treatments are alternated for some households and remain fixed for other households. Thus, period 2 serves as a control period for household 2 in

observing response in period 3, in addition to the usual "control" household number 1 which is never subjected to treatment. It is possible to subject different households to different durations of treatment to study the experimental duration effect. If these observations are spread over the business cycle, cyclical effects may be eliminated by averaging experimental response measurements over the business cycle.

As Sherwin Rosen has observed:

> We as a profession have engaged in excessive division of labor with regard to microdata collection. Thinking about survey instruments themselves and how they relate to economic phenomena and economic theories is probably an area where the social rate of return is fairly large.[35]

The sort of ongoing longitudinal data collection and experimental effort proposed here would encourage a wide range of research activities and give economists some private as well as social motivation for worrying about data collection and social experimentation.

Summary and Conclusions

We discussed some basic considerations involved in the evaluation of the methodology of social experiments. Many of the points raised seem obvious but, unfortunately, a number of them did not receive adequate attention in past social experiments in economics, for example in past negative income tax experiments. In our opinion, in most of these experiments, inadequate attention was given to formulating clear-cut attainable objectives. Feasibility studies and "test" or "pilot" experiments were nonexistent or not pursued vigorously enough. Serious measurement problems were encountered in these experiments and not dealt with adequately. Subject matter specialists, for example design statisticians, survey experts and outstanding subject matter theorists, were underrepresented or absent in the planning and execution of these experiments. Management and administration procedures were not completely satisfactory. The objectives of policymakers and of researchers usually were not clearly stated and in agreement. The experimental designs and the models on which they were based were inadequate in many cases. Last, the quality of reporting of results was generally far lower than could have been realized.

Some will say that the personal evaluations presented in the previous paragraph are "hypercritical" and that the negative income tax experiments constituted a valuable "learning experience." If so, this learning experience was very expensive and costly in terms of actual outlays and opportunity costs, including potential benefits associated

with successful social experiments and other uses of scarce research resources. If learning was a main objective, then it is doubtful that the design actually used to achieve this objective was a very good one, as Rosen (1985, p. 137) has stressed.

In the previous sections, we have attempted to provide constructive suggestions for improving the methodology of social experiments. Among the points made, these seem particularly important:

1. It is critical to design experiments for successful prediction of observable outcomes that are central to the objectives of an experiment and to provide useful measures of predictive accuracy, preferably complete predictive distributions. Sample sizes should be large enough to yield needed precision in prediction and the range of the design space should be large enough to attain the objectives of an experiment.

2. When there is uncertainty regarding appropriate models for the observations, experimental designs that provide information for discriminating among candidate models should be employed. In this connection, it has been recognized that many existing designs are very sensitive to model misspecification, for example errors in choice of functional form, departures from independence, and use of univariate models when multivariate models are more appropriate.

3. A mixture of model-based and randomized designs seems most appropriate, with carefully designed randomization procedures employed to guard against certain types of possible model misspecification and prejudicial elements. ANOVA-based designs are not adequate because they are very sensitive to model misspecification, they involve the need for many experimental units when a large number of extraneous variables have to be controlled and, most importantly, they are incapable of generating the predictions required in many social experiments.

4. Predictive validation of models used in social experiments is essential. For example, the labor supply equations estimated in the Seattle experiments can be employed to predict labor supply using data from the Denver experiment and vice versa. Unsatisfactory predictive performance is usually an indication of model misspecification, differential selection and other types of bias, poor data, or other flaws. Further, vigorous diagnostic checking of models in other ways, for example residual analyses and outlier detection procedures, is also recommended. Use of inadequate models vitally affects the internal and external validity of experiments.

5. Use of point estimates alone to appraise costs of alternative negative income tax programs, in the very few cases in which cost estimates were derived, is inadequate. Measures of precision or predictive probability distributions should be provided and interpreted in easily understandable terms for the benefit of policymakers. For exam-

ple, the probability that the costs of a program lie between $20 billion and $30 billion, or the probability that the costs exceed $30 billion, can be calculated and reported. Similar remarks apply to predictions of changes in hours of work. In both of these instances, it is the case that departures from independence of outcomes for experimental units can have an extremely large impact on precision measures, for example standard errors of total estimated costs and changes in hours. There was little attention given to these points in past social experiments in spite of the fact that such dependencies are of great concern in survey work, econometric analyses of panel data and past work on experimental design.

6. Consideration of dynamic theoretical labor supply models leads to models for observations that are radically different from the generally static models employed in most past social experiments, and their use would lead to different designs for experiments and different models for analyzing experimental data. It is recognized that the forms of such dynamic models are often uncertain and thus the use of unrestricted transfer function models, univariate or multivariate, may be a good point of departure in design and analysis calculations. Also, as stated above, design for discriminating among models can be effective in dealing with model uncertainties in dynamic as well as static cases.

7. It is recommended that, when feasible, social experiments be linked to ongoing longitudinal data generating programs of well-established groups, a suggestion put forward years ago by Orcutt and Orcutt (1968). With such an arrangement, historical variables have been measured that are useful in before-and-after calculations, as is done in "event" or "intervention" analyses. A longitudinal design also permits individuals to be used as their own controls. This is a standard technique in experimental designs in biology and psychology.[36] Longitudinal designs can provide improved results and deserve much further study. In particular, their use permits exploration of dynamic models and possible successful extrapolation of experimental results to a national population, given that the longitudinal sample is a national one. Of course, administrative costs and other aspects of national, longitudinal experimentation require attention.

While we have pointed to many difficulties involved in past social experiments, it is our opinion that *properly conducted* social experiments can yield enormous social benefits. Perhaps the objectives of past experiments have been too broad and ambitious, a point also made by Griliches (1985). Limiting objectives of social experiments in economics may be essential for attaining success. Successful experience with experimentation in the areas of experimental economics, quality management, marketing, and agricultural economics tends to support this view. Also, "on-line" experimentation to appraise proposed changes in existing social programs probably will be fertile ground for social experi-

menters who draw on the growing quality-management literature on this topic.

Finally, we have noted that the negative income tax experiments were focused on variants of the negative income tax proposal put forward many years ago by Friedman (1962), Tobin et al. (1967) and others. Unfortunately, the information provided by these experiments was not generally considered in relation to possible fundamental modifications of the original proposals. Among possible modifications, one might be the use of time paths for tax payments different from those used in the negative income tax experiments. To subject a poor person who begins to work to a marginal tax rate of 50 to 70 percent *immediately* is an extreme "treatment." It seems feasible to formulate a more sensible temporal pattern of tax payments that would avoid these high, initial marginal rates, a topic for future research and, perhaps, additional social experimentation.

This research was financed in part by the National Science Foundation and the H.G.B. Alexander Endowment Fund, Graduate School of Business, University of Chicago.

[1]With respect to the New Jersey negative income tax experiment, Rossi and Lyall (1976) conclude, "When it came down to the congressional debate on FAP, it was evident that while the labor supply question interested some congressmen in a general way, concerns were addressed more to the total costs of a national program, an issue to which the experiment could not offer an answer even when complete. It is one of the apparent ironies of the experiment that while its motivation sprang from a strong concern with poverty and a desire on the part of both the experimenters and OEO to effect national welfare reform, its most substantial contributions may well be of a more scholarly sort in the area of experimental design and work behavioral response." (pp.176-77.)

[2]Rossi and Lyall (1976) remark with respect to the New Jersey experiment, "Economists played dominant roles in all phases of the experiment . . . Sociologists and social psychologists were to play minor roles in both the design and analysis. Not only are the strengths of economists reflected in the experiment, but also some of the mistakes and omissions of the experiment show the mark of the dominant economists." (pp. 10-11.)

[3]See Friedman (1962, p. 193) and Tobin, Pechman and Mieszkowski (1967) for calculations of the costs of existing welfare programs and of negative income tax programs.

[4]Spiegelman and Yaeger (1980, pp. 474-476) provide a useful discussion of reporting error in the Seattle-Denver income maintenance experiments. On the basis of a "large sample wage income study," they report that, "SIME/DIME participants reported between $100 and $300 less per year to the experiment than to the Internal Revenue Service. This amount is less than 5 percent of mean income. The variance in the amount underreported to SIME/DIME is on the order of $1,000, or about one-fifth of mean income. We observed that almost as many people overreported their incomes as underreported them." To understand these and other measurement problems, they state that "Further study of individual cases is necessary." These conclusions relate to wage income, which is probably easier to measure than non-wage income. Ferber and Hirsch (1982) present much useful material on measurement problems in the negative income tax experiments.

[5]It is surprising to us that M. Friedman (1986) and J. Tobin (1986), two leading experts on negative income tax proposals, did not play major roles in the experiments. Tobin (undated) did provide some comments on the design of the New Jersey experiment. He wrote, "I find an "anova" specification implausible for this problem. But I recognize that

there is a certain arbitrariness to any particular parametric specification." (p. 18)

[6]When single equation regression or response surface models were employed for design purposes, possible dependence of observations was rarely, if ever, considered.

[7]See McFadden (1985) and Ferber and Hirsch (1982) for valuable considerations of this range of issues.

[8]For example, observation of the population at various stages of the business cycle is relevant for negative income tax experiments. Seasonality is also relevant.

[9]See Hausman and Wise (1979, 1985, p. 208) and Robins and West (1986) for analyses of attrition bias and efforts to deal with it. Robins and West (1986) conclude on the basis of their analysis of Seattle-Denver data that "Our results suggest that standard procedures of correcting for attrition bias do not always yield the proper results. The use of these procedures, however, depends to a large extent on the ability to model the attrition process and on the degree of attrition in the sample. In the SIME/DIME sample in which attrition was fairly modest . . . such techniques simply do not have the power to identify precisely the biases." (p. 337) In spite of these reservations, the authors conclude that "attrition bias is not a serious enough problem in the SIME/DIME data to warrant extensive correction procedures." A similar conclusion was reached by Hausman and Wise (1979) in their analysis of the Gary income maintenance experiment. (p. 937). Ferber and Hirsch (1982, p. 75 and p. 95) have reservations about such conclusions, however.

[10]See the appendix to Hamilton et al. (1969) for a discussion of the importance of good management in large scale research projects.

[11]Rossi and Lyall (1976) give the following breakdown of total costs of the New Jersey negative income tax experiment.

A. Administration and Research		
Mathematica	$4,426,858	
IRP-U. of Wisconsin	812,648	
sub total		$5,239,506
B. Transfer Payments		2,375,189
C. Grand Total		$7,614,695

They state, "The expenditures were a considerable overrun on the initial estimates of approximately $3 million. Most of the unanticipated expenses occurred on the research side. The handling of large and complicated data sets was simply much more costly than anyone anticipated." (p. 11) These comments underline the importance of good management techniques in the planning and execution of social experiments. For further discussion of these issues, see Ferber and Hirsch (1982).

[12]See footnote 1 for a possible illustration of this point in connection with the N. J. experiment.

[13]See Hausman and Wise (1985) for an exellent collection of articles on key aspects of analysis of economic experiments and Aigner and Morris (1979) for extensive discussion of designs of these experiments.

[14]See Watts and Bawden (1978) for discussion. Of course, some social experiments such as the Housing Supply Experiment are not feasible without a site approach. It would be impossible to discern a supply effect without involving a large percentage of households in particular housing markets.

[15]See Keeley et al. (1978a) for an example of this sort of calculation for a national negative income tax. Labor supply response functions were coupled with census household data in a "micro-simulation" of the national program.

[16]The experiences with changes in the New Jersey welfare laws during the course of the experiment highlight the importance of diversification. State-to-state differences in program implementation are to be expected in a national implementation of a negative income tax program. An experiment based on only one or two states cannot possibly take into account variation in local welfare programs.

[17]Conlisk and Watts (1979), p. 40.

[18]See Morris (1979) for a discussion of the finite selection model which essentially provides a randomization technique for providing more balanced experimental designs. The finite selection model utilizes the same sort of objective function and cost constraint as the Conlisk/Watts model.

[19]Hausman and Wise (1985) and Keeley et al. (1980) point out that current designs do not ensure that statistically significant treatment effects can be obtained.

[20]Keeley et al. (1978, p. 11).

[21]It was not until 1978 that Keeley et al. produced a thorough analysis of costs of a

national negative income tax based on labor supply response estimates from experimental data.

[22]See Keeley et al. (1978b, Table 3, p. 13) for estimates.

[23]The fitted model is a truncated normal regression or tobit model. The interpretation of σ as the standard deviation of prediction error conditional on knowing the model parameters is strictly not correct. σ should be interpreted as the standard deviation of prediction error for the latent variable. At levels of the independent variables for which little truncation occurs it is approximately correct to view σ as the root mean squared error of prediction.

[24]See Stafford (1985) for discussion of these studies and a table of elasticity estimates.

[25]Diffuse priors may be used in studies with little or unreliable nonexperimental data.

[26]See Hausman and Wise (1979a,b), Heckman (1978) and Hausman and Wise (1985) for discussion of modeling approaches to these problems.

[27]See Rossi (1985) for details of these calculations and an application to choice between alternative functional forms.

[28]Heckman (1976) makes this point.

[29]See Duan et al. (1984). This model has been employed earlier in econometrics by Orcutt, Goldberger and many others.

[30]See Zellner and Rossi (1984) for an example of this approach for binomial response models and Zellner (1971, 1984).

[31]Griliches (1985), p. 138.

[32]The Seattle-Denver income maintenance experiment contained treatments of three, five and 20 years. It is very difficult to find discussion of the results for 20-year treatments.

[33]Within the context of linear models, these dynamic specifications yield restricted transfer function models.

[34]The survey and income dynamics survey conducted at University of Michigan and the national longitudinal labor survey are examples.

[35]Rosen (1985, p. 137).

[36]Rossi and Lyall (1976), p. 42, fn. 24. See also Campbell and Stanley (1963), and Hall (1975).

References

Aigner, Dennis J. "A Brief Introduction to the Methodology of Optimal Experimental Design," *Journal of Econometrics*, 11 (1979a), pp. 7-26.

_____. "Sample Design for Electricity Pricing Experiments: Anticipated Precision for a Time-of-Day Pricing Experiment," *Journal of Econometrics*, 11 (1979b), pp. 195-205

Aigner, Dennis J., and C.N. Morris, eds. *Experimental Design in Econometrics, Supplement to the Journal of Econometrics*, 11 (1979), pp. 1-205.

Box, G.E.P., and W.J. Hill. "Discrimination Among Mechanistic Models, *Technometrics*, 9 (1967), pp. 57-71.

Box, G.E.P., and G.C. Tiao. "Intervention Analysis with Applications to Economic and Environmental Applications," *Journal of the American Statistical Association*, 70 (1985), pp. 70-79.

Burtless, Gary, and Jerry A. Hausman. "The Effect of Taxation on Labor Supply: Evaluating the Gary Negative Income Tax Experiment," *Journal of Political Economy*, 86 (1978), pp. 1103-1130.

Burtless, Gary, and Larry L. Orr. "Are Classical Experiments Needed for Manpower Policy?" unpublished manuscript, 1986.

Cain, Glen G., and Harold W. Watts, eds. *Income Maintenance and Labor Supply*, Chicago: Markham, 1973.

Campbell, D.J., and J.C. Stanley. "Experimental and Quasi-Experimental Designs for Research on Teaching," in N.L. Gage, ed., *Handbook of Research on Teaching*, Chicago: Rand McNally & Co., 1963, reprinted in book form, *Experimental and Quasi-Experimental Designs for Research*, Chicago: Rand McNally & Co., 1973 (tenth printing).

Conlisk, John. "Choice of Response Functional Form in Designing Subsidy Experiments," *Econometrica*, 41 (1973), pp. 643-656.

_____. "Design for Simultaneous Equations," *Journal of Econometrics*, 11 (1979), pp. 63-76.

_____. "Comment", in Jerry A. Hausman and David A. Wise, eds., *Social Experimentation*, Chicago: U. of Chicago Press 1985, pp. 208-214.

Conlisk, John, and M. Kurz. "The Assignment Model of the Seattle and Denver Income Maintenance Experiments," Res. Mem. #15, Center for the Study of Welfare Policy, Stanford Research Institute, 1972.

Conlisk, John and Harold W. Watts. "A Model for Optimizing Experimental Designs for Estimating Response Surfaces," *Journal of Econometrics*, 11 (1979), pp. 27-42.

Covey-Crump, P.A.K., and S.D. Silvey. "Optimal Regression Designs With Previous Observations," *Biometrika*, 57 (1970), pp. 551-566.

Crutchfield, J.A., and Arnold Zellner. *Economic Aspects of the Pacific Halibut Fishery*, Washington, DC: U.S. Department of the Interior, Government Printing Office, 1963.

Duan, N., W.G. Manning, Jr., C.N. Morris, and J.P. Newhouse. "Choosing Between the Sample-Selection Model and the Multi-Part Model," *Journal of Business and Economic Statistics*, 3 (1984), pp. 283-289.

Ferber, Robert and Werner Z. Hirsch. "Social Experiments in Economics," *Journal of Econometrics*, 11 (1979), pp. 77-115.

_____. *Social Experimentation and Economic Policy*, Cambridge: Cambridge U. Press, 1982.

Fienberg, S.E., B. Singer and J.M. Tanur. "Large-Scale Social Experimentation in the United States," in A.C. Aitkinson and S.E. Fienberg, eds., *A Celebration of Statistics: The ISI Centenary Volume*, New York: Springer-Verlag 1985, pp. 287-326.

Fisher, Ronald A. *Statistical Methods for Research Workers* (1st ed.), New York: Hafner Publishing Co., 1925.

Friedman, Milton. *Capitalism and Freedom*, Chicago: U. of Chicago Press, 1962.

_____. personal communication, 1986.

Geisel, M.S. "Bayesian Comparisons of Simple Macroeconomic Models," in S.E. Fienberg and A. Zellner, eds., *Studies in Bayesian Econometrics and Statistics in Honor of Leonard J. Savage*, Amsterdam: North-Holland Publishing Co., 1975, pp. 227-256.

Griliches, Zvi. "Comment," in Jerry A. Hausman and D.A. Wise, eds., *Social Experimentation*, Chicago: U. of Chicago Press, 1985, pp. 137-138.

Grossman, J.B. "Optimal Sample Designs With Preliminary Tests of Significance," *Journal of Business and Economic Statistics*, 4 (1986), pp. 171-176.

Guttman, I. "A Remark on the Optimal Regression Designs with Previous Observations of Covey-Crump and Silvey," *Biometrika*, 58 (1971), pp. 683-685.

Hall, Robert E. "Effects of the Experimental Negative Income Tax on Labor Supply," in Joseph A. Pechman and P. Michael Timpane, eds., *Work Incentives and Income Guarantees: The New Jersey Negative Income Tax Experiment*, Washington, DC: The Brookings Institution 1975, pp. 115-156.

Hamilton, H.R., S.E. Goldstone, J.W. Milliman, A.L. Pugh, E.R. Roberts, and A. Zellner. *Systems Simulation for Regional Analysis: An Application to River-Basin Planning*, Cambridge: MIT Press, 1969.

Hausman, Jerry A., and David A. Wise. "Attrition Bias in Experimental and Panel Data: The Gary Income Maintenance Experiment," *Econometrica*, 47 (1979a), pp. 455-474.

_____. "Social Experimentation, Truncated Distributions and Efficient Estimation," *Econometrica*, 45 (1979b), pp. 919-938.

_____, eds. *Social Experimentation*, Chicago: U. of Chicago Press, 1985.

Heckman, James. "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement*, 5 (1976), pp. 475-492.

_____. "Sample Selection Bias as a Specification Error," *Econometrica*, 47 (1979), pp. 153-161.

Jeffreys, H. *Theory of Probability*, London: Oxford U. Press, 1967.

Keeley, Michael C. *Labor Supply and Public Policy: A Critical View*, New York: Academic Press, 1981.

Keeley, Michael C., Philip K. Robins, Robert G. Spiegelman, and Richard W. West. "The Estimation of Labor Supply Models Using Experimental Data," *American Economic Review*, 68 (1978a), pp. 873-887.

_____. "The Labor Supply Effects and Costs of Alternative Negative Income Tax Programs," *Journal of Human Resources*, 13 (1978b), pp. 3-36.

Keeley, Michael C., and Philip K. Robins. "Experimental Design, The Conlisk-Watts Assignment Model, and the Proper Estimation of Behavioral Response," *Journal of Human Resources*, 15 (1980a), pp. 480-469.

_____. "The Design of Social Experiments: A Critique of the Conlisk-Watts Model and Its Application to the Seattle-Denver Income Maintenance Experiments," in R.G.

Ehrenberg, ed., *Research in Labor Economics* (Vol. 3), Greenwich, CT: JAI Press, 1980b, pp. 293-333.
Kempthorne, O. *The Design and Analysis of Experiments,* New York: John Wiley & Sons, Inc., 1952.
MacCrae, E.C. "Optimal Experimental Design for Dynamic Economic Models," *Annals of Economics and Social Measurement,* 6 (1977), pp. 379-405.
McFadden, D.L. "Comment," in J.A. Hausman and D.A. Wise eds., *Social Experimentation,* Chicago: U. of Chicago Press, 1985, pp. 214-219.
Metcalf, Charles E. "Making Inferences from Controlled Income Maintenance Experiments," *American Economic Review,* 63 (June 1973), pp. 478-483.
_____. "Predicting the Effects of Permanent Programs from a Limited Duration Experiment," *Journal of Human Resources,* 9 (1974), pp. 530-555.
Morris, C. "A Finite Selection Model for Experimental Design of the Health Insurance Study," *Journal of Econometrics,* 11 (1979), pp. 43-61.
O'Hagan, A. "Curve Fitting and Optimal Design for Prediction," *Journal of the Royal Statistical Society B,* 40 (1978), 1-42 (with discussion).
Orcutt, G.H., and A.G. Orcutt. "Incentive and Disincentive Experimentation for Income Maintenance Policy Purposes," *American Economic Review,* 58 (1968), pp. 754-772.
Palmer, John L. and Joseph A. Pechman, eds. *Welfare in Rural Areas: The North Carolina-Iowa Income Maintenance Experiment,* Washington, DC: The Brookings Institution, 1978.
Pechman, Joseph A., and P. Michael Timpane, eds. *Work Incentives and Income Guarantees: The New Jersey Negative Income Tax Experiment,* Washington, DC: The Brookings Institution, 1975.
Peck, S.C., and R.G. Richels. "On the Value of Information to the Acidic Deposition Debates," ms., Electric Power Research Institute, Palo Alto, CA, forthcoming in *Journal of Business and Economic Statistics.*
Rivlin, Alice. "Allocating Resources for Policy Research: How Can Experiments Be More Useful?" *American Economic Review Papers and Proceedings,* 64 (1974), pp. 346-354.
Robins, Philip K., and Richard W. West. "Labor Supply Response Over Time," *Journal of Human Resources,* 15 (1980), pp. 525-544.
_____. "Sample Attrition and Labor Supply Response in Experimental Panel Data," *Journal of Business and Economic Statistics,* 4 (1986), pp. 329-338.
Rosen, S. "Comment," in J.A. Hausman and D.A. Wise, eds., *Social Experimentation,* Chicago: U. of Chicago Press, 1985, pp. 134-137.
Rossi, Peter E. "Comparison of Alternative Functional Forms in Production," *Journal of Econometrics,* 30 (1985), pp. 345-361.
Rossi, Peter H. "A Critical Review of the Analysis of Nonlabor Force Responses," in Joseph A. Pechman and P. Michael Timpane eds., *Work Incentives and Income Guarantees: The New Jersey Negative Income Tax Experiment,* Washington, DC: The Brookings Institution, 1975, pp. 157-182.
Rossi, Peter H. and K.C. Lyall. *Reforming Public Welfare: A Critique of the Negative Income Tax Experiment,* New York: Russell Sage Foundation, 1976.
Rubin, D.B. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology,* 66 (1974), pp. 688-701.
Spiegelman, Robert G., and K.E. Yaeger. "Overview," *Journal of Human Resources,* 15 (1980), pp. 463-479.
Stafford, F.P. "Income-Maintenance Policy and Work Effort: Learning from Experiments and Labor Market Studies," in Jerry A. Hausman and D.A. Wise, eds., *Social Experimentation,* Chicago: U. of Chicago Press, 1985, pp. 95-134.
Tobin, James. "Sample Design for NIT experiment," memo to Harold Watts and William Baumol, 19 pp, undated.
_____. personal communication, 1986.
Tobin, James, Joseph A. Pechman, and Peter N. Miezkowski. "Is a Negative Income Tax Practical?" *Yale Law Journal,* 77 (1967), 1-27, reprinted in James Tobin, *Essays in Economics,* Cambridge, MA: MIT Press.
Watts, Harold W. and D.L. Bawden. "Issues and Lessons of Experimental Design," in John L. Palmer and Joseph A. Pechman, eds., *Welfare in Rural Areas: The North Carolina-Iowa Income Maintenance Experiment,* Washington, DC: The Brookings Institution, 1978.
Zellner, Arnold. *An Introduction to Bayesian Inference in Econometrics,* New York: John Wiley & Sons, Inc., 1971.
_____. *Basic Issues in Econometrics,* Chicago: U. of Chicago Press, 1984.
Zellner, Arnold, and Peter E. Rossi. "Bayesian Analysis of Dichotomous Quantal Response Models," *Journal of Econometrics,* 25 (1984), pp. 365-393.

Discussion

*Jerry A. Hausman**

After a brief introduction Arnold Zellner and Peter E. Rossi turn to relevant considerations for evaluation of social experiment methodology. They discuss eight considerations which encompass design, management, and analysis of social experiments. In general their list provides a good common sense approach to the subject. I would like to stress their seventh point: that experiments should be designed to provide "accurate enough" predictions of various proposed policies along with measures of predictive precision. Because of the large amount of inherent variability in responses to tax and welfare policies, even within the assumption of a correctly specified model to evaluate the responses, two aspects of the Zellner-Rossi prescription should be emphasized. First, the range of policies that the experimental results will be used to evaluate must be specified with sufficient precision so that the experiment covers them. Otherwise, extrapolation outside the range of the experiment will be required, with undesirable consequences. This goal is often very difficult to achieve without increasing the costs greatly, and this aspect of design is especially dependent on the "specialists" Zellner and Rossi refer to. Second, the design and results must be able to supply results that are sufficiently precise to use. My greatest disappointment with the negative income tax experiments has been the low level of precision that arose from the results. Future social experiments should make sufficient precision in outcomes among their highest priorities in design.

Zellner and Rossi turn next to design considerations within a static framework. They criticize the Conlisk-Watts design for the negative income tax experiments for too restricted variation in experimental treat-

*Professor of Economics, Massachusetts Institute of Technology.

ment. Since I have discussed the Conlisk-Watts design elsewhere (Hausman 1982 and Hausman and Wise 1985), I will not return to previous ground. However, I would like to point out that the Zellner-Rossi criticism holds only within the context of a structural model of labor supply, for example the famous Elfving result for linear models, which places all the observations at the extreme points of the design space. In an ANOVA framework the response to each treatment point is estimated separately, so the Zellner-Rossi criticism does not apply. Even within the context of a structural model, I have considerable doubt whether I would want to use the responses. Our structural models are not usually sufficiently well specified that they can do a good job on extreme points in the sample space.

Next in their discussion of the Conlisk-Watts approach, Zellner and Rossi emphasize the specification of a demand equation for allocations of subjects across treatment groups. However, they fail to take into account the complexity of the actual demand equations that arise in response to government programs. For instance, the labor supply equation (leisure demand equation) will not be continuous even in the tax rates because of the nonconvexities in the budget sets. (See Burtless and Hausman 1978 and Hausman 1985.) The "housing gap" treatment in the Housing Allowance Demand Experiment has similar characteristics. This very complicated response surface is quite different from the response surfaces in many physical situations, where the responses are apt to be smooth. Zellner and Rossi should consider in more detail the complications for experimental design which these types of demand functions imply. These complications should induce a more favorable attitude to randomization procedures, which Zellner and Rossi discuss but do not strongly advocate.

Zellner and Rossi emphasize that a decision-theoretic approach would more likely lead to results that are usable by policymakers. While they stress a Bayesian approach to the problem of reporting results, I have found that an analogous "classical" approach, with point estimates and standard errors that account for parameter uncertainty, are straightforward to compute and seemingly well understood by public utility commission staffs who have evaluated results from experiments. I am in total agreement with Zellner and Rossi that the results of an experiment should be sufficiently precise to yield predictions with enough precision to give good guidance to policy. As Zellner and Rossi demonstrate in their analysis of the Keeley et al. results (1978) from the Seattle-Denver experiments, the negative income tax experimental designs do not lead to precise predictions about the labor supply response, which was certainly one of the major goals of the experiments. (Note that the Zellner-Rossi estimates of the prediction error of the Keeley results would be considerably larger if parameter uncertainty

were accounted for, since this uncertainty is correlated across all observations in a microsimulation.)

Zellner and Rossi then turn to dynamic aspects of social experiments. They emphasize correctly that the experiments typically are of short duration, while the policies are permanent in nature and may therefore call forth a different response. However, I disagree with their suggestion that a Box-Jenkins times series approach would be a useful starting point for analysis. Lagged endogenous variables are quite difficult to treat in short panels because of initial condition problems; more importantly, the errors of measurement, which Zellner and Rossi emphasize earlier in their paper, have potentially devastating effects on times series type models of panel data. (See Griliches and Hausman 1986.) I do agree with their suggestions on the usefulness of panel data, which I discuss with respect to social experiments in Hausman (1982). However, it must be noted that panel data may raise the costs considerably for an experiment because of the necessity of keeping track of panel members. The cost trade-off between panel data and cross-section data would need to be considered, as Heckman has emphasized in recent research.

Zellner and Rossi conclude that the goal, design, execution, and analysis of the negative income tax experiments left much to be desired. I agree with these conclusions in large part. However, I believe their failings can be partly explained by the design and execution of the Gary and Seattle-Denver experiments before the lessons of the New Jersey experiment were learned. Presumably better experiments would be conducted now. My major point of disagreement lies in the analysis of the data: I believe that Zellner and Rossi have too much faith in structural models and that their time series approach to longitudinal data would not work well. But, we certainly agree that such experiments should be designed so as to be able to answer the important questions at issue in a precise enough manner to be useful for planning and policy purposes.

References

Burtless, Gary, and Jerry A. Hausman. "The Effect of Taxation on Labor Supply: Evaluating the Gary Negative Income Tax Experiment," *Journal of Political Economy*, 86 (1978), pp. 1103-1130.

Griliches, Zvi and Jerry A. Hausman. "Errors in Variables in Panel Data," *Journal of Econometrics*, 1986.

Hausman, Jerry A. "The Effect of Time on Economic Experiments," in W. Hildebrand, ed., *Advances in Econometrics*, Cambridge University Press, 1982.

_____. "The Effect of Taxes on Labor Supply," in Alan J. Auerbach and Martin Feldstein, eds., *Handbook of Public Economics*, 1986.

Hausman, Jerry A., and David A. Wise, eds., *Social Experimentation*, Chicago: University of Chicago Press, 1985.

Keeley, Michael C., Philip K. Robins, Robert G. Spiegelman, and Richard W. West. "The Estimation of Labor Supply Models Using Experimental Data," *American Economic Review*, 68 (1978a), pp. 873-887.

_____. "The Labor Supply Effects and Costs of Alternative Negative Income Tax Programs," *Journal of Human Resources*, 13 (1978b), pp. 3-36.

Discussion

*Charles E. Metcalf**

Arnold Zellner and Peter E. Rossi review the conventional criticisms of the methodology of the early income maintenance experiments—which by now have had 19 years to accumulate—and provide their own suggestions for design of social experiments. Unfortunately, the authors' own recommendations fare poorly against the standards of criticism applied to previous work, and show insufficient evidence of the 19 years of experience that have accumulated since the design work for the first negative income tax experiment began. My comments follow the approximate sequence of the paper.

Considerations for Evaluating Methodology

In the first part of their paper, Zellner and Rossi offer eight considerations for evaluating the methodology of a social experiment. These conventional observations are largely correct but naively elaborated upon. For example, the call for "interaction between sponsors and bidders in the preparation of proposals" reflects a simplistic view of the competitive procurement process, but does touch on an important issue: the complex relations among contractor selection, project design, and project execution. Indeed, it is increasingly common for the design and execution phases of an experiment or evaluation to be the subjects of separate contract procurements.

Concerning the desirability of conducting a "pilot" trial of an experiment before proceeding with the "final" experiment, the distinc-

*President, Mathematica Policy Research, Inc. Views expressed are the sole responsibility of the author.

tion between these concepts is blurred in an evaluation environment cluttered with an extensive history of social experiments and demonstrations. We must also keep in mind that each "desirable" characteristic of an experiment has an opportunity cost, not the least of which is the passage of time. (Most suggestions for improving methodology tend to increase the duration of an experiment.) While many people—myself included—view the social experiments as having made important contributions to the policy process, provision of timely input with respect to originally specified experimental objectives is rarely one of them.

Failure to acknowledge opportunity costs also causes the authors to overstate another observation, which carries forward to their critique of the negative income tax experiments: "If the objectives . . . involve generalization . . . to an entire population, then the sample . . . *has to be* a sample from the relevant population." (Emphasis added.) It is equally true, however, that the program intervention being tested has to be the "relevant" intervention—in terms of features of program administration, duration, and so forth—and these two objectives are frequently in conflict. An experimental design stressing intervention with the right population is not clearly preferable to an experiment that restricts the population to improve the intervention.

Static Design Issues

Several static design issues raised by Zellner and Rossi are worthy of comment. First, the claim that the planners of the negative income tax experiments exaggerated the problems with administration and field operations of a national experiment is probably true for data collection, but *not*, in my judgment, for program administration. Recall that an effective implementation of a program intervention requires—aside from its placement in an effective evaluation structure—creation of a "relevant" program environment as viewed by the experimental subject; real and perceived independence of program administration from data collection; and at least some participation of welfare agencies in all jurisdictions covered by the experiment. These pressures all work to limit the number of jurisdictions covered by an experiment, and are further enhanced by the increasing prevalence of the view that "relevant" program interventions must be implemented by "real" program agencies rather than by experimenters, in order to be credible. This evolution is paralleled by a clear transition from experiments that test parameters to randomized demonstrations that test program interventions. There has been a recent trend toward the use of representative samples for demonstrations and/or experiments, but with cluster samples utilizing "real" program interventions in a relatively few sites.

Second, the authors criticize the negative income tax experiments for being too conservative in their choice of design parameters for the experiments. From a pure design perspective most experimenters would agree with the authors. But policymakers with whom the experimenters had to interact were reluctant to consider the concept of "extreme" experimental treatments outside the "policy-relevant range."

Third, the authors provide an extensive discussion contrasting the "response surface" and ANOVA approaches to design, and stressing importance of the analytic models that drive the experimental design. I agree with much of the authors' position here, and my disagreements with them are more often of form than of substance. Several points, however, are worth raising:

- The response surface approach is described as producing "nonrandomized" designs. This is true of the finite selection model extreme, but not of the Conlisk/Watts approach, which determines probabilities of selection for each element in the design space. So long as no probabilities of selection are permitted to go to zero, there exists an ANOVA equivalent for each response surface design.
- The potential damage caused by use of an inappropriate design model depends upon whether its use eliminates design points called for by the "correct" model, or whether it merely reduces estimation efficiency for the correct model. A linear or Cobb-Douglas model would spell disaster for the estimation of a translog function, but the converse is not so.
- In the (universal?) case where the correct model is not known with certainty, a risk-averse design strategy involves use of a model with more "dimensions" than specific models likely to be investigated, preferably with all probabilities of selection constrained to be positive. Inclusion of an ANOVA model as one of several weighted alternatives fulfills this objective. In such an environment it would not be surprising for the full design model never to be used for analysis.

Fourth, I do not regard the role of controls in social experiments as being "unusual" in their use of the status quo rather than the classical "no treatment" as the basis of comparison. The control group should reflect a relevant counterfactual, which may or may not meet the semantic definition of "no treatment." Consider also that removal from previously existing treatment is "no treatment" in only the most unrealistic of static worlds. As for whether controls are necessary except as cheap observations, this depends upon the experimental objective. For most policy purposes, as well as most reasonable predictive procedures, the relevant counterfactual is a critical component of evalua-

tion. Indeed, I would regard the proper objective of the negative income tax experiments *not* to be estimation of the national cost of a negative income tax for comparison with external cost data for AFDC; rather, they should be providing internally valid *direct estimates* of the *differential cost*. I would argue this point on both policy and statistical grounds.

Fifth, I do not regard the discussion of cross-unit dependence as being particularly relevant from an empirical perspective, since rho = .01 is *massive* when applied for each unit to each of 40,000,000 other units, not "small" as alleged by the authors. If I were looking for a reason to disregard nominal standard errors obtained from the experiments, I would make a simple appeal to cluster sampling theory. For similar reasons I would not use labor supply functions fitted to Denver data to predict response for the Seattle sample, as suggested by the authors for validation, since the relevant sample size is *two* in too many dimensions. Rather, I would recommend a traditional split sample approach cutting across site boundaries to achieve that objective.

Dynamic Design Issues

The authors' discussion of dynamic design issues goes rather smoothly until they take seriously the notion of a longitudinal panel as the basis for drawing experimental samples, which takes the flawed concept of letting individuals be their own controls to an unfortunate extreme.

Concerning their general discussion, I would be careful to distinguish between two important but separate issues: the use of *limited-duration interventions* in place of relevant longer-term interventions (for example, the negative income tax experiments) and *limited-duration observation* of longer-term dynamic consequences. Time series models, for example, deal with the latter but not the former problem.

I have no quarrel with advocacy of better longitudinal data sources for continuing evaluation and research, often as an *alternative* to randomized experiments. The development of the SIPP panel appears to be especially promising. On the other hand, evidence is mounting that efforts to use longitudinal panels as comparison group alternatives to randomized control groups have been unsuccessful.

Similarly, the theoretical concept of an experimental panel has merit so long as it can provide an adequate sample, so long as the relevant program interventions can be applied to it, and so long as the sample points are disposable rather than reusable. Sample adequacy is a major problem, since many program interventions of policy interest are targeted to relatively small segments of the population. Earlier in my comments I questioned the ability to create the relevant program en-

vironment with a dispersed sample for most social programs of the sort earmarked for experiments or demonstrations.

Finally, the concept of reusing sample points in repeated experiments sounds fine when all interventions and impacts are static, but in a world of dynamic interventions and impacts the cross-experimental contamination effects would appear to destroy all credibility of the experimental results. Continuing panels for data collection are fine; for controlled interventions, extremely questionable.