Volume Title: Social Experimentation

Volume Author/Editor: Jerry A. Hausman and David A. Wise, eds.

Volume Publisher: University of Chicago Press

Volume ISBN: 0-226-31940-7

Volume URL: http://www.nber.org/books/haus85-1

Publication Date: 1985

Chapter Title: Introduction to "Social Experimentation"

Chapter Author: Jerry A. Hausman, David A. Wise

Chapter URL: http://www.nber.org/chapters/c8371

Chapter pages in book: (p. 1 - 10)

Introduction

Jerry A. Hausman and Davis A. Wise

During the past decade the United States government has spent over 500 million dollars on social experiments. The experiments attempt to determine the potential effect of a policy option by trying it out on a group of subjects, some of whom are randomly assigned to a treatment group and are the recipients of the proposed policy, while others are assigned to a control group. The difference in the outcomes for the two groups is the estimated effect of the policy option. This approach is an alternative to making judgments about the effect of the proposed policy from inferences based on observational (survey) data, but without the advantages of randomization. While a few social experiments have been conducted in the past, this development is a relatively new approach to the evaluation of the effect of proposed government policies. Much of the $500 million has gone into transfer payments to the experimental subjects, most of whom have benefited from the experiments. But the most important question is whether the experiments have been successful in their primary goal of providing precise estimates of the effects of a proposed government policy. This book is a collection of papers and comments from a conference held in 1981 under the sponsorship of the National Bureau of Economic Research and supported by the Alfred Sloan Foundation. At the conference papers were presented that addressed the question of the success of the experiments in achieving this evaluation goal.

In addition to the question of the success of the experiments for policy evaluation, whether the experiments were worth the cost was a recurring question among the conference participants. That is, could similar information have been provided by the use of econometric models and estimates on available survey data? It is important to remember that the policies that were evaluated in many of the experiments were far different from then-current policies that could be evaluated from survey data. For

example the income-guarantee levels for male-headed households in the negative-income-tax experiments were far higher than any state provided in its welfare system. Similarly, time-of-use electricity prices for residential customers were unheard of in the United States before the experiments began. Nevertheless, estimated income elasticities from models of labor supply based on, say, the Survey of Economic Opportunity could be used to predict the outcome of the negative-income-tax experiments, and data from Europe might be used to predict the effect of the time-of-use electricity experiments. The authors of each of the first four papers that evaluate what we have learned from the four major groups of experiments focus on this question as well as the measured effect of the experiments.

It is important to keep in mind the conclusions based on hindsight, and they should be evaluated in this light. Even if the results of the experiments could have been well predicted by previous econometric estimates, no one could have been confident of this outcome before the experiments actually took place. Indeed this was a major motivation for the experiments. The authors of the four evaluation papers also consider what other purposes the results of the experiments are used for because experimental results may be superior to econometric estimates for these other uses. But since, to a large extent, the policy questions the experiments were designed to answer still have not been decided, the final accounting of the worth of the experiments in helping to decide the course of public policy is probably fairly far off into the future.

The methodology of randomized experiments was formalized and achieved wide acceptance due to the fundamental research of R. A. Fisher and his co-workers. Yet most of their research dealt with agricultural experiments. Complex statistical questions arise when this methodology is applied to social experiments. Individuals may refuse to participate in experiments or they may drop out while the experiment is underway. These effects must be accounted for in the evaluation of the experimental results. The next set of papers addresses questions of statistical and econometric methodology with respect to social experiments. Questions of sample selection and treatment assignment are analyzed. Also an overall model of how the results will be used is postulated. The possibility exists of a decision-theoretic approach to the evaluation of policies which could lead to a considerably sharper focus in the design of the experiments.

The last group of papers takes up the extremely complicated question of how the output of the experiments is actually used in policy formulation. Experiments provide new information that can lead to reevaluation of previously proposed policies or the formulation of new policy options. The analysis of the experiments and the policy-formulation process become intertwined, and yet the latter is constrained to some extent by the

political process. What are the conditions under which the analysis of the experiments has the most influence on the policy process? Again the final impact of the experiments will not be known until more time passes, but these questions are important to the design and analysis of successful social experiments.

We now turn from this general introduction to a summary of each of the papers presented at the conference. Each paper was commented on by either one or two discussants, and we also summarize their comments. The authors had the opportunity to revise their papers after the conference and in some instances reacted to discussants suggestions. But most of the comments deal with broad questions motivated by or raised in the papers and raise valuable additional points about the success of social experiments and possible improvements in their implementation.

Dennis J. Aigner in his paper, "The Residential Electricity Time-of-Use Pricing Experiments: What Have We Learned?" evaluates the fifteen time-of-use (TOU) experiments that have taken place since 1975. The main question at issue is whether TOU prices would produce alterations in the demands of residential customers for peak-period electricity large enough so that the net effect on the customer's welfare plus the change in revenues and investments required of the electric utility would justify implementation of TOU rates. That is, the TOU rates are a close application of marginal-cost pricing by the utility. The experiments were designed to discover whether the change in rate pattern to a time-of-day basis would lead to an increase in social welfare. In his summary of the results, Aigner finds much less agreement among the price-elasticity estimates than is found in the labor-supply studies of the NIT experiments. However, he does find that the results lead to the conclusion that peak-period demands are inelastic when expenditure is held constant, which can have important implications for the effects of a TOU rate plan on utility revenues. Aigner then goes on to consider the welfare effects of the TOU rate plans. Based on the results of the experiments he concludes that only the largest customers are likely to benefit from the introduction of TOU rates. However, these large customers consume a significant proportion of total residential electricity demand. Lastly, Aigner takes up the difficult question of "transferability." Can the results of an experiment in one jurisdiction be used to predict the outcome of a TOU plan in another area? Aigner offers an interesting approach to this important problem.

Overall, Aigner concludes that only a very few experiments have led to reliable results. However, he does not find the differences in elasticity estimates too disturbing when considered across different service areas. His view is that better experimental design would have led to considerably better results. He also points out a potentially important limiting factor of the experiments. The experiments are best analyzed as short-run

experiments since the appliance stock has been held fixed. He believes that the long-run response could be considerably different from the response suggested by estimates based on the experimental results.

Paul L. Joskow, in his comments, agrees with Aigner's conclusion on the limited usefulness of many of the TOU experiments. However, he doubts that the neoclassical welfare analyses used by Aigner to evaluate the TOU rates will be acceptable to state regulatory commissions. He also emphasizes that the short-run nature of the experiments will limit their usefulness. Therefore, he concludes that the TOU rate experiments will have little if any positive impact on regulatory decisions to implement TOU rates. Lester D. Taylor, in his comments on the Aigner paper, concludes that the evidence is not good enough to make a scientific evaluation of the desirability of TOU rate plans. Thus, all three authors feel that we have not "learned enough" from the TOU rate experiments although they all conclude that a limited amount of knowledge has been gained through the experiments.

Harvey S. Rosen reviews the housing-allowance experiment in his paper, "Housing Behavior and the Experimental Housing-Allowance Program: What Have We Learned?" This experiment granted housing subsidies to the poor to determine to what extent they would increase their housing consumption. The Experimental Housing Allowance Program (EHAP) was divided into two parts. The demand experiment was designed to determine the effect on housing consumption of the subsidies. The supply experiment was designed to determine the effect of the housing allowances on the rental housing market. Given the relatively inelastic short-run supply of rental housing units, it is important to estimate the effect of a subsidy program on market rents. Rosen analyzes the value of EHAP in terms of what additional knowledge we gained from the experiment which could not have been known from previous econometric studies on cross-sectional data. He bases his criterion on the argument that a structural econometric model is necessary to analyze housing consumption patterns, even with experimental data. He considers the problems that arise in econometric estimation of housing-demand functions, but he concludes that these same problems were present in the analysis of the EHAP demand data so that the experiment did not alleviate the usual problems that applied work in housing-demand encounters. His conclusions are quite similar for the problems that exist in the analysis of the EHAP supply experiment. Thus, overall Rosen argues that the problems faced by investigators who have used conventional survey data continue to exist in the experimental data except for variations in the price of housing induced by the experiments. He does not think that the social experiment was necessary and concludes "The money would have been better spent on augmenting conventional data sources."

John M. Quigley broadly agrees with Rosen's conclusions. He outlines some more complete analytical models for problems that arise in the housing experiment. He raises the additional problem that the duration of the EHAP demand experiment was quite short, since it lasted for only four years. Therefore inferences about the long-run response may be problematic. He feels that the effect of long-term subsidies might be quite different from the observed response to the EHAP subsidies. Gregory K. Ingram, in his discussion of the Rosen paper, also comes to similar conclusions about the value of EHAP. He does think, however, that program-participation rates which are an important determinant of program costs would be difficult to predict without an experiment. But his overall assessment is that EHAP did not help solve the many problems of measurement that exist in the analysis of housing markets. He concludes that the EHAP did have some value, but at too high a cost. He believes that only the demand experiment of EHAP was worthwhile.

In his paper, "Income-Maintenance Policy and Work Effort: Learning from Experiments and Labor-Market Studies," Frank P. Stafford reviews the evidence from the largest and perhaps most important group of experiments, the negative-income-tax (NIT) experiments. The five NIT experiments were designed primarily to analyze the effects of a potentially large change in the income support or welfare system in the United States. Since the cost of an NIT program would be closely related to the labor-supply response of individuals to the income guarantee and the tax rate, these parameters were varied across individuals or families in the experimental design. The response of individuals or families to the introduction of an NIT in terms of their work effort is closely related to their labor-supply behavior. Stafford's first question is "Why did we need the experiments at all?" He argues that from previous studies on survey data, labor economists had formed a consensus view on the range within which the labor elasticities would fall. Therefore estimates of the effect of the introduction of an NIT could be made from the coefficient estimates from these previous studies. The case for the NIT experiments from this vantage point he believes is perhaps not overwhelming.

Stafford discusses two reasons for the NIT experiments. One possible role is that the experiments would be easier to understand and to interpret by policy makers who would place more confidence in their results than in simulation estimates from survey data. The NIT experiments provide direct evidence on the alteration of work effort which could well be more convincing. The second reason for the NIT experiments, which Stafford finds less convincing, is that "model free" analysis of the experiments is possible while the related labor-supply studies must be based on econometric models of questionable validity. This latter reason is less convincing, in Stafford's viewpoint, because any actual NIT plan likely to be adopted is unlikely to be exactly one of the experimental treatments,

and a model will be necessary to predict its labor-supply effects. Whether or not one finds the case for theoretical models to be strong, Stafford concludes that a strong case exists for the experiments to answer the main question of the effect of an NIT on work effort.

Besides the effect of an NIT on work effort, other outcomes of interest which Stafford identifies are on-the-job training, divorce or change in family structure, and labor-market outcomes of unemployment, work effort, and early retirement. In terms of the main variable of interest, work effort, Stafford finds the results of the NIT experiments broadly consistent with nonexperimental studies. Stafford argues that to answer the question of whether the experiments were "worth it" would require a decision theoretic model that could evaluate the possible policy alternatives and take account of the greater precision, or less uncertainty, that would result from the experimental evidence. Evidence from the NIT experiments on the other areas of labor-market behavior is valuable, but not conclusive, in Stafford's opinion. The effect of the experiments on greater divorce rates points up the important effects of transfer systems on family decision making. Overall, Stafford concludes that a great deal was learned from the experiments. Furthermore, he sees the possibility of continued research using the data that was collected, which would help answer other important questions.

Sherwin Rosen, in his comments, emphasizes the decrease in maintained model hypotheses which an experiment allows. He emphasizes that the finding of a similar work-effort response found in the NIT experiments as is found in survey data could not have been known in advance. While he thinks that room for improvement in design and analysis of the experiments certainly exists, overall he concludes that the NIT experiments led to a valuable increase in our knowledge in the area of work response to change in income guarantees and tax rates. In his comments, Zvi Griliches emphasizes the importance of randomization in experiments and the "exogeneity" introduced by experimental treatments. Griliches also emphasizes the importance of the experiments in providing increased variation in the data which in general will lead to better econometric estimates. Overall, he takes a somewhat stronger view than does Stafford on the value of the experiments.

Jeffrey E. Harris analyzes the health experiments in his paper, "Macroexperiments versus Microexperiments for Health Policy." His major point is that health experiments may be better designed and analyzed at the community or group level which makes them differ fundamentally from the microexperiments at the individual level of most other types of economic and social experiments. Harris first considers the problems inherent in the microexperiments that have been conducted so far. He concentrates on the Multiple Risk Factor Intervention Trial (MRFIT) and the Rand Health Insurance Study (HIS). He claims that data from

MRFIT are difficult to analyze because of the problem of participation bias. He also has doubts about the sample selection procedure in the HIS experiment. He then considers the potential problems of attrition bias, of interdependence among individual responses, and of Hawthorne effects. Overall, he thinks that these problems lead to quite complicated model designs to analyze the effects of the experiments. Harris feels that many of these problems could be minimized by the use of macroexperiments such as the Stanford Heart Disease Prevention Program (SHDPP) which uses as the unit of observation a complete community. While he does not find the SHDPP without fault, he believes that its main problem could be alleviated by a different experimental design. He argues that repeated cross-sectional sampling in macroexperiments is the preferred design. He favors the macroexperiments mainly because use of the mass media becomes a valued policy option since it does not affect the behavior of controls that are geographically distinct communities. He feels that further development of the theory of the design and analysis of macroexperiments would be useful since they offer the opportunity of more convincing experiments than do the microexperiments with their insurmountable difficulties.

Paul B. Ginsburg has reservations about Harris's claims for macroexperiments. He feels that the use of macroexperiments is limited by cost considerations. He thinks that many experimental situations have elements of both micro- and macroexperiments so that the choice between the two types is not often clear-cut. In his comments, Lawrence L. Orr takes sharp issue with Harris's conclusions. Orr thinks that a careful analysis is needed to decide which type of experiment is more useful in a particular situation. He disagrees most fundamentally with Harris over the question of whether the role of an experiment is to decide on the efficacy of a particular policy option or whether it is to analyze a range of possible policies. He believes that the latter situation is more typical so that microexperiments are needed to estimate individual response functions. The "black box" macroexperiment then becomes inappropriate. Furthermore, he believes that many of the problems inherent in microexperiments also exist in macroexperiments. Orr also believes that additional important problems of interpretation exist in the results of macroexperiments. Most important of these limitations is the sample-size constraint in macroexperiments together with the difficulty of control and administration. Orr concludes that the particular situation must be analyzed to determine whether a micro or a macro approach is more appropriate. He differs strongly with Harris's conclusion on the superiority of macro- over microexperiments.

The next set of papers considers the question of experimental design and analysis. Jerry A. Hausman and David A. Wise in "Technical Problems in Social Experimentation: Cost versus Ease of Analysis" attempt to

set forth general guidelines that would enhance the usefulness of future social experiments and to suggest methods of correcting for inherent limitation in the experiments. They feel that more attention should be paid to the possibility of randomized design and its associated analysis. The experiments to date have utilized endogenous sample-selection procedures and treatment-assignment procedures which subvert the possibility of using classical analysis-of-variance procedures to determine the results of the experiments. Still, inherent limitations exist, even with randomized design, which are difficult or impossible to avoid. The problems of voluntary participation and of attrition from the experiment will continue to exist. But Hausman and Wise argue that these problems are considerably easier to treat if the confounding problems of endogenous stratification and assignment are not present. Lastly, they propose that the experiments be designed to estimate only a small number of treatment effects rather than a large range of policy options which often leads to imprecise estimates of the effect of any single policy option.

John Conlisk agrees with Hausman and Wise that endogenous stratification should be avoided if possible. However, he is in less agreement with the principle of random treatment assignment. He thinks that interactions between treatments and exogenous variables are of central importance to the behavioral response of the experiments. He also thinks that the number of experimental design points must be based on particular design considerations so that an overall judgment cannot be readily made. He concludes that self-selection and attrition problems are of great importance and that experimental design theory needs to be extended to deal with these problems. In his comments on the Hausman-Wise paper, Daniel L. McFadden concurs with the use of robust techniques such as ANOVA for the analysis of the experiments. He argues that random treatment assignment is the most important factor in an acceptable design in that it isolates the effects of other sample-frame difficulties. Endogenous sampling designs can then be used, although the statistical analysis is complicated somewhat. He agrees that problems of self-selection and attrition are important and suggests that the focus of future research should be on robust statistical methods to correct these problems.

Frederick Mosteller and Milton C. Weinstein consider the cost effectiveness of the experiments in their paper, "Toward Evaluating the Cost-Effectiveness of Medical and Social Experiments." Mosteller and Weinstein emphasize the importance of learning about the efficacy and cost-effectiveness of medical practices so that decisions about proper medical procedures can be made. They then proceed to consider the costs and benefits of the evaluation procedures. Therefore, they propose to evaluate the evaluations. They evaluate the procedure of a randomized clinical trial (RCT). To do so they consider the question of how the evaluations are actually used. They then specify a general conceptual

model for evaluating the cost-effectiveness of the clinical trials. They do so by using a decision analytic model to assess the cost-effectiveness. They formulate a Bayesian model and after deriving the results consider the effect of relaxing some of the assumption of the model.

Mosteller and Weinstein then consider inherent problems in the assessment of cost-effectiveness of medical evaluations. They discuss both normative and positive models of response to the evaluations. The question of institutional design, to assure appropriate use of the information, is also covered. Other problems such as the assessment of information in the experiments and their utilization are discussed.

Then Mosteller and Weinstein turn to examples to examine the usefulness of their suggested approach. The examples deal with the gastric-freezing procedure and the treatment of hypertension. They conclude that further study is needed but that a potential cost-benefit calculation should be made before a trial is undertaken. Furthermore, they argue that controlled experiments cannot remove all the problems of evaluation but that they are of value and should be utilized more often. They believe that an incentive system which would lead to increased tests of efficacy would be advantageous policy.

In his comments Joseph B. Kadane agrees with the Mosteller and Weinstein conclusions about the need for more evaluation. He points out some of the limitations of their model and questions how the potential cost-benefit calculation that Mosteller and Weinstein call for could be made without additional information.

In "The Use of Information in the Policy Process: Are Social-Policy Experiments Worthwhile?" David S. Mundel argues that social-policy experiments are very expensive and therefore should be undertaken only in very particular situations. He argues that the potential utility of social experiments depends on the following factors: whether the experiments can answer the questions that are important to policy makers; whether the answers can be understood, given that important policy questions can be answered; and finally, whether the answers alter the beliefs of policy makers, given that they provide understandable answers to important questions.

Ernst W. Stromsdorfer examines the effect on policy of social experiments in his paper, "Social Science Analysis and the Formulation of Public Policy: Illustrations of What the President "Knows" and How He Comes to "Know" It." Stromsdorfer begins with the contention that policy makers will use whatever data are at hand to support their position whether or not the data come from a social experiment. He therefore considers the larger questions of how information is used in the policy-formulation process and how it interacts with the political process in the making of policy decisions. Stromsdorfer sees three processes for knowledge production in the federal government: management information-

system (MIS) data, natural or quasi experiments, and classical experiments. He concludes that while a variety of experimental data exists of the natural, quasi-experimental, and classical experimental variety, they are of uneven quality and are not used in a consistent manner.

But Stromsdorfer does believe that this information has been used in the consideration of policy issues in recent Congressional and administration decisions. He points to the issues of welfare reform which have been affected by the results of the NIT experiments, unemployment insurance which has been affected by numerous studies for natural or quasi experiments, and social security reform. At the same time evaluation research often follows the lead of policy development. But Stromsdorfer also identified research categories that have had little or no impact. Overall, he concludes that program analysis and evaluation can be an extremely valuable policy tool. But at the same time, the reality of political constraint must be recognized since it sets limits on the collection, analysis, and usefulness of data and analysis which may be produced by experimentation.

Henry Aaron in his comments on Stromdorfer's paper agrees with the focus of research being used in an adversary process. Furthermore, Aaron emphasizes that the adversaries are contending for power, not the scientific truth that might arise from the experiments. But Aaron concludes that social experiments have been a force for slowing the adoption of new policies. Social experiments show problems to be more complicated than is commonly appreciated, with results more difficult to achieve. Lawrence E. Lynn, Jr., in his comments takes no view on whether the experiments have been useful. He warns against the overuse of the "rational actor" model of the political process.