SOCIAL EXPERIMENTATION: SOME WHYS AND HOWS

PREPARED UNDER A GRANT FROM THE U.S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE

RAE W. ARCHIBALD, JOSEPH P. NEWHOUSE

R-2479-HEW
MAY 1980

Rand
SANTA MONICA, CA. 90406

SOCIAL EXPERIMENTATION: SOME WHYS AND HOWS

PREPARED UNDER A GRANT FROM THE U.S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE

RAE W. ARCHIBALD, JOSEPH P. NEWHOUSE

R-2479-HEW
MAY 1980

Rand
SANTA MONICA, CA. 90406

PREFACE

This report is one product of Rand's Health Insurance Study, a social experiment sponsored by the Department of Health, Education, and Welfare. A portion of the research program, including this report, consists of using Rand's experience to draw lessons about the technology of social experimentation.

The report will also appear as a chapter in the *Handbook on Systems Analysis*, to be published by the International Institute for Applied Systems Analysis, Laxenburg, Austria. The report's intended audience consists of people who are to make decisions on implementing social experiments, and people who will either monitor the progress of or actually conduct social experiments. Section I is addressed primarily to the former group, Secs. II and III to the latter.

Because of its purpose, the report presents little new material on social experimentation. Exactly which research projects should be classified as social experiments is to some degree a matter of judgment. Such matters as whether randomization is a necessary characteristic and how results might best be generalized to a larger population are among the issues beginning to be explored in the small but growing literature of social experimentation. This report reflects the authors' interpretation of views dominant in the literature, but the reader should be aware that alternative views exist.

The authors believe, however, that the existing literature does not adequately treat topics related to the management of an experiment. Considerable space has been devoted to that subject here in the belief that management is critical to the success or failure of an experiment.

SUMMARY

This report consists of three parts: when to conduct a social experiment (and when not to), how to manage a social experiment, and some practical advice ("tips") to experimenters. Social experimentation compares people who receive a certain program (experimental treatment) with similar people who do not or who receive some other treatment. Some form of randomization is typically used to maximize the likelihood that the various groups are in fact similar.

WHEN TO EXPERIMENT

The experimental method is the best available for establishing that a certain program or intervention actually caused a given result or set of results. Other methods often encounter difficult problems. Typically, they compare groups that are (or may be) different and attempt to adjust for differences; if the adjustment is not fully satisfactory, it may leave the cause-and-effect relationship ambiguous. Furthermore, other methods often have to work with data already in existence, which may be difficult to analyze and which may not bear on the proposed program or treatment, or can be brought to bear only by making unverifiable assumptions. When such problems prevail, inferences about program effects may be questionable.

Obstacles may also impede or even preclude social experimentation. Perhaps most important, it is usually more expensive and time-consuming to collect and analyze experimental data than it is to use existing data; consequently, the experimenter should be sure that existing data will not suffice for the purpose before deciding to go ahead with an experiment—and should also be sure that the problem to be studied is important enough to warrant the expenditure of resources. Problems that could vitiate an experiment include: insufficient observations (usually a subcase of the experiment's being too expensive for the problem), inability to define or control the treatment variable, and important outcomes that are not measurable.

HOW TO MANAGE AN EXPERIMENT

Because of its unique aspects, the considerable attention given to social experimentation in the literature has surrounded the manager of an experiment with an exaggerated mystique—exaggerated because the manager must possess, above all, the traditional managerial skills needed for planning, operations, financial management, personnel management, public relations, and communications.

Experiment personnel are likely to be organized into a research team or teams and an action or operations team or teams (including a data processing group). Because it will be responsible for analyzing the data, the research team must have ultimate authority over the experiment, but early and frequent interaction between the teams is necessary for success. The teams must respect each other and

understand the constraints imposed on each other. One of the manager's crucial duties is to preserve the integrity of the treatments, a task for which it is essential to create information systems that permit monitoring the field operations.

Planning and cost estimation are both important skills; deficiencies in them can undermine the best designed experiments. Most experiments will have highly interrelated field operations; failure to consider dependencies among them is virtually certain to lead to poor-quality data. Accurate cost estimation is essential if the experiment has a fixed budget. If costs prove higher than estimated, necessary ad hoc changes in the design of the experiment may cause much more damage than would have occurred if the higher costs had been foreseen. If costs prove lower than estimated, opportunities to collect useful information may have been missed.

The data processing group should participate in the early stages of design if its operations are to be efficient and effective. Most social experiments generate a large data base, and unless the data are gathered with an eye toward the methods by which they will ultimately be processed, the research team may well find itself able to use only a portion of the data.

PRACTICAL ADVICE TO THE EXPERIMENTER

Several "tips" can be given to the prospective experimenter: Create a pilot sample of people with whom to pretest the operational feasibility of the experiment. Build into the design an ability to measure effects that are an artifact of the experiment ("methods" effects or Hawthorne effects). Construct the experiment to keep refusal and attrition at low levels, especially refusal and attrition correlated with treatments. Usually, balance the sample across treatments. Do not strongly oversample a group whose membership is not well defined. Go to some lengths to inform participants about the treatment and their obligations. Choose the number of sites and length of enrollment to minimize variance, given a fixed project budget. Do not attempt too much, but do not be easily discouraged.

ACKNOWLEDGMENTS

The authors received helpful comments on an earlier draft from Henry Aaron, Gene Fisher, James Gaither, Thomas Glennan, David Novick, Edward Quade, Henry Riecken, Noralou Roos, and Peter Szanton. Naturally, these individuals should not be held responsible for any remaining shortcomings. The views expressed in this paper represent those of the authors and not necessarily those of the sponsors of Rand research.

CONTENTS

I. INTRODUCTION

Social experimentation should occupy an important place in the system analyst's toolkit. Riecken et al. (1974) offer the standard definition of a social experiment:

> An experiment is one or more treatments (programs), representing intervention into normal social processes, that are administered to some set of persons (or other units) drawn at random from a specified population; observations or measurements are made to learn how (or how much) some relevant aspects of their behavior differ from those of a group receiving either another treatment or no treatment, also drawn from the same population.

Social experimentation is sometimes likened to the laboratory experiment in the natural sciences. But because the social experimenter has somewhat less control than the laboratory scientist normally has, a better analogy is with the clinical trial in medicine (Cochran, 1972).

In comparison with other methods, the great advantage of an experiment is that from it, when it is properly designed and executed, the analyst can infer most confidently that a given intervention (program) actually caused a given result (Riecken et al., 1974; Gilbert, Light, and Mosteller, 1975).

In many situations, however, the analyst is called in after the fact, to work with data already collected by some experimental program or treatment in which the analyst may have had little or no say in the selection of sample groups. For example, the program may have been administered to those who applied for it first, or who were judged to be in greatest need; the analyst's task may be to infer what effect the program might have if extended to the general population. The analyst may not even have had much say over what data were collected.

Such situations are termed quasi-experimental or demonstration projects; appropriate research protocols for these situations are described by Campbell and Stanley (1966). We do not discuss such protocols here, although we do describe why a true experiment will usually be more informative.

AN EXAMPLE OF AN EXPERIMENT'S POTENTIAL

A nonexperimental study done by Kessner et al. (1973) illustrates the value of an experiment in making causal inferences. These researchers studied infant mortality among mothers who did and did not receive adequate prenatal care, according to the researchers' criteria. Controlling for the mothers' socioeconomic status, the mortality rate among children whose mothers received adequate care was found to be 16 percent below the rate among children whose mothers did not.

On its face, that result suggests that a large-scale prenatal care program could substantially lower the infant mortality rate. Yet one would be rash to draw this inference, because mothers who received adequate care probably differed system-

atically from those who did not—even controlling for socioeconomic status. For example, one criterion for adequate prenatal care was that the mother voluntarily sought care in the first trimester. It is likely that mothers who did so were also more motivated to exercise and keep their weight down than were mothers who did not. In such a case, a difference in infant mortality rates might well have been observed between the two groups of mothers even if they had all received adequate care. In other words, some of the difference may have been due not to prenatal care but to characteristics of the mothers.

How much a prenatal care program for all mothers might reduce infant mortality rates cannot be ascertained using Kessner's methods, but a properly conducted social experiment could yield the answer. What would the experiment look like? The new program could be instituted with a randomly chosen group of mothers. If, after a period of time, the infant mortality rate differed between that group and a similar group who did not receive the program (but could, of course, seek care on their own), the program could be instituted generally. If differences did not appear (and the analysis were precise enough to rule out all but negligible differences), a more effective program design could be attempted.

If one chose not to conduct an experiment, the alternatives would then be: (1) not to institute the program at all; (2) to institute it universally; (3) to institute it partially, but not on a randomized basis. Let us assume that enough is known to make the potential program appear promising; in particular, the odds that the program will successfully reduce the infant mortality rate justify the resources that an experiment would require. Otherwise, the program should probably not be instituted at all.[1]

To institute the program universally, one must be confident that the program will "work"; but it may be infeasible to do so for administrative or budgetary reasons. If so, an experiment is almost certainly preferable to nonrandom partial introduction, because it will provide evidence on whether the program should be continued.[2] Even if one could implement the program universally, it is often preferable to evaluate it experimentally, and that requires nonuniversal implementation.

In this report we consider the decision to undertake an experiment, its organization, and some possible pitfalls in designing it. The discussion is based on the authors' experience at Rand with designing and managing a social experiment, the Health Insurance Study (HIS), which began in 1971 and the field work for which will end in 1982. (For a description of the study's experimental design, see Newhouse, 1974.) We do not describe basic experimental design, because many texts and monographs are available.[3]

---

[1]Justification for introducing the program in spite of its high cost would have to rest on assumed improvement in nonmeasurable outcomes, e.g., social cohesion. If this were the case, an experiment of the type described would not be useful, as pointed out below.

[2]If it is argued that some mothers are clearly more in need, it may be possible not to randomize the neediest of mothers and still conduct an experiment.

[3]Aigner, 1979; Campbell and Stanley, 1966; Cochran and Cox, 1957; Cook and Campbell, 1974; Cox, 1958; Federov, 1972; John, 1971; John and Quenouille, 1978; Kempthorne, 1952; Kendall and Stuart, 1968; Scheffe, 1959; Srivastava, 1975. Discussions of social experimentation, including many examples of experiments not discussed in this paper, may be found in Boruch and Riecken, 1975; Boruch et al., 1978; Ferber and Hirsch, 1978; Orr, 1974; Orr, Hollister, and Lefkowitz, 1971; Plott, 1979; Riecken, 1977; Riecken and Boruch, 1978; Rivlin, 1974; and Wilson, 1974, in addition to the other references herein. Boruch (1974) presents a list of illustrative experiments.

THE ADVANTAGES OF SOCIAL EXPERIMENTATION

Determination of Causality

The most important advantage of social experimentation has already been identified: A properly designed and executed experiment can provide the strongest evidence that certain programs or policy actions actually cause or, if implemented, would cause certain outcomes. Because the experimenter controls the program policy or treatment and who receives it, one can be more confident than with other methods that an association between program and outcome was not spurious. (In statistical terms, the program or policy is definitely exogenous.) Several social experiments have been undertaken because existing data did not permit analysts to determine causal relationships with sufficient confidence. Recent examples include studies of income maintenance and Rand's HIS study.

Income Maintenance. A variety of experiments in income maintenance have been conducted in the United States.[4] In these experiments, families were guaranteed varying levels of income, even if they earned nothing. Payments to the family were reduced ("taxed") as family income rose. After a certain income level was reached, payments ceased altogether. At issue was the degree to which families might reduce their hours of work as support levels (guarantees) and alternative tax rates varied. Would workers, for example, quit their jobs and simply live off the guaranteed income?

The economics literature contains numerous nonexperimental studies of this matter (e.g., Cain and Watts, 1973). These studies infer how families might respond to proposed income maintenance programs by analyzing how much work is performed by people with differing wage and property incomes. Such an analysis might show, for example, that workers earning $6 an hour worked 5 percent less than workers earning $8 an hour. One might then hypothesize that if the tax rate on the $8-an-hour worker were raised enough to reduce after-tax wages to $6 an hour, he or she would work 5 percent less. Unfortunately, one requires a strong assumption to reach that conclusion. People who prize material goods may both invest in training that yields a high wage rate and then willingly work long hours. People who believe that the best things in life are free may prize their leisure time more than the rewards of work. It would be rash to assume, then, that high-wage workers would slacken their efforts if their wage were lower, or that higher wages would inspire the others to work more. Nonetheless, nonexperimental studies had to make this assumption, whereas the income maintenance experiments induced variation in the actual tax rate and studied the number of hours that people worked.[5]

The Health Insurance Study. In the Health Insurance Study (HIS), one objective was to measure how the use of medical care varied with the completeness of insurance coverage (e.g., how much use would increase if medical services were

---

[4]Bawden and Harrar, 1976; Keeley, et al., 1978a, 1978b; Kehrer, 1977; Kershaw, 1972; Kershaw and Fair, 1976; Palmer and Pechman, 1978; Pechman and Timpane, 1975; Rossi and Lyall, 1977; Watts and Rees, 1977.

[5]Technically, the taste for leisure was kept independent of variation in after-tax wage rates by varying the rate at which the guarantee was taxed away and by randomly assigning individuals to different tax rates. There was exogenous variation in "after-tax" wage rates.

free). A nonexperimental method for achieving this objective would be to compare, say, the numbers of visits to physicians made by people who do and who do not have complete health insurance. One trouble with that measure, however, is that people in the United States who have more complete insurance rate themselves less healthy than do those with less complete insurance (Phelps, 1976). Because the two groups therefore differ in more ways than their insurance coverage, a comparison of the two could overstate the responsiveness of use to insurance. Statistical methods have been designed to correct for such bias (i.e., simultaneous equation methods), but in this instance they yielded highly imprecise estimates of the effect of insurance (Newhouse and Phelps, 1976). In the experiment, however, similar groups could be assigned to each insurance plan; consequently, one could be reasonably confident that any difference in the use of medical services across plans was attributable to the insurance.[6]

Social experimentation offers a number of other advantages over analysis of nonexperimental data; these include: expanding the range of variation in existing data; making analysis of some problems more tractable; ensuring data availability; and providing evidence on the administrability of a proposed program.

Expanding The Range of Variation in Existing Data

Existing data are often insufficient for evaluating a proposed program because they do not apply to the relevant range.

In the income maintenance experiments, a key issue was how much low-income households would reduce their work effort as the size of the guaranteed income rose.[7] An estimate can be derived from nonexperimental data by observing how hours of work vary among households with differing amounts of property income. (Property income is economically analogous to a guarantee.) But low-income households rarely have much property income, and almost none had as much as the income maintenance program proposed to give them as a guarantee. Hence, estimates using nonexperimental data required extrapolation of an estimated relationship well outside its observed range.

In the peak-load electricity pricing experiment reported by Manning, Mitchell, and Acton (1976), an estimate was desired of how much households would shift their electricity consumption from peak-hours to off-peak hours if electricity cost more during peak-hours. If some consumption shifted to off-peak hours (for example, if households used washing machines only at night), utility capacity costs could be reduced. No American data addressed this issue, because peak-load pricing had never been tried in the United States; that is, a kilowatt-hour cost the same around the clock. Some European data did exist, but because of different conditions (for example, differences in climate and in stocks of household appliances), it was decided to undertake an experiment in the United States.

Of course, the necessary data for an experiment may not exist because the proposed experimental treatment is infeasible or undesirable. In such cases the

---

[6]Technically, the insurance plan is exogenous and, to the degree possible with a finite sample, orthogonal to other covariates.

[7]Technically, what was desired was an estimate of the income effect on labor supply among low-income households.

experimental information can have little value. Before mounting an experiment, then, the experimenter needs to ask why data do not exist—that is, why existing institutions do not provide the program or treatment.

But a treatment might not be feasible at a universal level and still yield useful information, because it may show whether moving in a certain direction has any desirable effect at all. For example, it may be impractical to triple the number of foot patrolmen in a police system, but one may conduct a limited experiment to see if tripling has any measurable effect on crime at all. If it does not, one can dismiss the idea of increasing the number of foot patrolmen by a lesser amount.

Improving Analytical Tractability

Sometimes data exist in a form difficult to analyze, or the particular data necessary have not been collected. In such a case, the assumptions needed to estimate the effects of a program or policy can raise suspicions about the conclusions. Starting from scratch is therefore a great advantage: One can design the experiment and specify the data to be collected in ways that render the experimental results relatively easy to analyze. An example is provided by the Health Insurance Study.

In the United States, many health insurance policies vary the fraction of the bill reimbursed according to how many dollars have already been spent. One example is an initial deductible; the insured pays for the cost of medical care up to a fixed amount, beyond which insurance benefits begin. Policies that vary the fraction of expenditures that are reimbursed are difficult to analyze (Keeler, Newhouse, and Phelps, 1977). The experimental insurance policy was designed to avoid these analytical problems as much as possible by keeping the fraction of expenditures reimbursed constant except for families with infrequent, very large expenditures.

It might be argued that designing the experiment to allow simple analysis limits generalizability—for example, that experimental findings cannot be used to predict behavior under real-world insurance policies. (This generic argument has been made by Roos, 1975.) Although in a strict sense this objection is correct, it ignores the principle of divide and conquer. Analysis in this particular experiment examines (among other things) how the use of medical care services responds to rather large variations in how much the patient must pay (net of reimbursement) for medical services. If analysis of the experimental data were to show that use responded very little to changes in the extent of insurance (i.e., that people tend to consume about the same amount of medical care regardless of their insurance coverage), one could generalize to real-world policies with some confidence. Because the variation in the real-world policies is almost certainly smaller than variation in the experimental plans, it can be inferred that the variation in real-world policies has little effect upon demand.

Alternatively, the experimental results may show that insurance coverage makes a great deal of difference in the amount of medical care a policyholder consumes. In this case the analyst who wishes to estimate use for a real-world policy must decide which experimental policy best approximates the real-world policy.[8]

[8]Technically, the analyst must translate the real-world policy into the price space spanned by the experiment.

Although this endeavor must necessarily be imprecise, the experimental results could substantially reduce uncertainty about the effects of different insurance plans. If not, further experimentation with different types of policies might be appropriate.

Availability of Data

The analyst may not have the data necessary to make sound predictions of a proposed program's effects; new data therefore must be sought to enable an empirical analysis prior to implementing a program. For example, the Health Insurance Study sought to relate people's health status to their insurance coverage. Because no existing data set permitted a comprehensive analysis of this issue, the choice was between running an experiment or collecting data by observing households with whatever insurance policies they held. In either event new data had to be collected; in this case, the additional cost of an experiment may not be very large, and the experiment will offer all the other advantages discussed here.

Implementation Issues

When a new program is being contemplated, an experiment can sometimes reveal unsuspected implementation issues. Again, the income maintenance experiments serve as an example. Although a number of proposals had been made for a guaranteed income in the United States (e.g., Friedman, 1962; National Commission on Technology, Automation, and Economic Progress, 1965; The President's Commission on Income Maintenance Programs, 1969), little or no thought had been given to the demarcation of income-accounting periods at the time the first income maintenance experiment began (1968).

When the experiment was on the drawing boards, an issue arose concerning the time period over which income should be measured (accounting period). On the one hand, if the payments to the family were to be a function of the previous year's income, the family could be left in considerable need if, near the beginning of the present year, its income fell suddenly, but it qualified only for low payments because of its previously high income. On the other hand, if payments were to be a function of income over a shorter period of time, say a month or a quarter, normal seasonal fluctuations in the family's income (for which the family could plan) could inflate the family's payments. (An example of seasonal fluctuation is the situation of a farm family that receives its income when it sells its crop.) Not until serious experimental design began was it realized that alternative rules for income-accounting periods could have important consequences for cost and behavior.

An alternative to an experiment that will yield information about feasibility and administrability is a demonstration project. In such a project, one simply launches a program without attempting to obtain data about a comparable group of people who do not receive the program. But a demonstration project does not offer the other advantages of an experiment. In particular, it cannot answer the question of what difference the intervention made, and should probably not be used

unless the sole issue is feasibility. Even then, an experiment might cost little more than did the demonstration project.

WHEN NOT TO EXPERIMENT

Problems that could preclude an experiment include: adequacy of existing data; inability to obtain sufficient observations; inability to define or control a relevant treatment variable; unresolvable ethical issues; and important outcomes that are not measurable.

Adequacy of Existing Data

If existing data are adequate, an experiment should not be undertaken. The experiment will be much more expensive and time-consuming than analysis of existing data. It should therefore be undertaken only when existing data do not permit satisfactory answers and the importance of the issue justifies the resources that an experiment will require.

Insufficient Observations

The simplest experiments to implement are those that vary the program or treatment across individual families. The experimenter probably will have good control over the treatment, and the number of observations will usually suffice to provide the desired information. The experiments described above are of this type. By contrast, if the unit of observation is an entire geographic area, cost considerations may permit only a few observations and interfere with reliable estimation.

For example, several proposals have recently been made in the United States to restructure health insurance, with the intent of strengthening price competition in the medical care market (Ellwood, 1978; Enthoven, 1978; Newhouse and Taylor, 1971). These proposals share the feature that the insured's premium goes up if he or she chooses doctors or hospitals that charge a high price for their services or deliver a great many services. (In the United States, as in most countries, the insured patient typically bears little, if any, of the cost differences among doctors or hospitals that the patient might use.) The advocates of these proposals hope that consumers will take costs into account when choosing a doctor, thereby creating incentives for efficient production. They also hope the schemes will not cause access problems for people in chronic ill health, although they potentially could if doctors are rated by the billings they make to the insurance company.

Consider an experiment with a plan that varied insurance premiums with the doctor used. Most people in a given local medical market would have to be covered by the plan; otherwise, medical providers would have no incentive to give much attention to the prices they charged or quantity of services delivered. Thus, an experiment would require that the plan be instituted in some markets but not in other, comparable ones. But it may not be feasible or practical to introduce an experiment into many local markets; millions of people might have to be enrolled. Alternatively, one could perhaps implement the program in two or three localities.

Although such a small sample probably would not yield very precise estimates of the effects of a national program, one might well garner valuable information (especially about administrative problems), thereby permitting a more informed choice about the desirability of these proposals. In that case, however, the project would more resemble a demonstration than an experiment.

Inability to Define or Control the Treatment Variables

In some cases the experimenter can only partially specify the treatment or program; those administering the treatment to participants will also importantly shape it. Sometimes, for example, it may be advisable to allow considerable discretion to local program managers in adapting the program to their locality. This situation can pose a dilemma for the experimenter.

On the one hand, the experimenter may want full control, and could specify the treatment in such detail as to deny discretion to the local manager. On the other hand, if the experimenter wants to know what will happen if local managers have discretion, he or she may have to surrender a considerable degree of control. (Rigid control is not always possible, anyway.) The delegation of control also has its pitfalls. If the experimenter leaves the treatment poorly or incompletely specified to those implementing it, the results of the "experiment" may be uninterpretable and unreplicable, because no one knows what actually was implemented.

This issue did not arise to any appreciable degree in the income maintenance, health insurance, or peak-load pricing experiments described above. The treatment was well defined, and reasonably similar application in a national program could be assumed. In the income maintenance experiments, for example, the issue was whether the guarantees and tax rates were as the experimenter claimed (i.e., whether the experimenter maintained control of the treatment).

But when the ultimate program does not "reasonably" resemble the experimental treatment, the experiment will lose its value (Guttentag, 1973, 1977). Sometimes this is due to poor experimental management; we take up management issues in the next section. Sometimes the problem is discretion in implementation at the local level. If the proposed program includes an element of local discretion, the experimenter could implement the program in a group of randomly chosen localities or organizations (allowing local administrators leeway) and compare those localities or organizations with a control group. Such a design can be difficult to carry out; there may be too few areas for a reliable comparison, or residents of areas without the program may obtain its benefits by moving.

Problems of this sort would occur if one were to experiment with unrestricted revenue transfers from one level of government to another—for example, from the federal government to states. If the experimenter chose recipient states at random, probably little could be gained from comparing them with states that did not receive such revenues. The recipient states may have done very different things with the funds (i.e., implemented quite different programs or treatments), and any particular state program could have had numerous and different effects. The path between a decision to make unrestricted transfers to certain states and their ultimate effects is too long to make experimentation a very useful tool for evaluating

such transfers. Nonexperimental efforts to understand how the lower levels of government spent any additional resources afforded them seem more appropriate.

Inability to Resolve Ethical Issues

Ethics should seldom bar social experimentation (see Rivlin and Timpane, 1975). In general, we are speaking about experimenting with programs that are being seriously proposed as public policy. If such a program is unethical, then it should not be enacted into law, much less undertaken as an experiment. In the case of clinical trials in medicine, the experimenter must determine whether the known benefits of a treatment are so clear as to make it unethical to *withhold* the treatment. Rarely will enough be known about a proposed social program to make it unethical to withhold the treatment; even if it is thought that enough is known, the program possibly cannot be implemented immediately for the entire population anyway.

Nonexperimental studies also encounter ethical problems, although they are often largely unrecognized. Gilbert, Light, and Mosteller (1975) point out that experiments frequently show that a treatment had no effect (or perhaps even a deleterious effect) whereas nonexperimental evidence had indicated that the program would be beneficial. Errors in the other direction may also occur. If the nonexperimental studies are numerous enough, it may be difficult on ethical grounds to justify a controlled experiment. For example, if a nonexperimental study has concluded that a certain program is harmful, it is possible to argue that a controlled experiment would be unethical. Rightly or wrongly, then, part of the price of a nonexperimental study is the decreased likelihood of later experimentation.

Although ethics will seldom preclude an experiment, they should shape the experimental design. Ethics dictate that the design minimize risk to participants; reduction of risk will also lower refusal and attrition, thereby increasing the value of the data collected. Some techniques to reduce risk are described in more detail in Sec. III below; they include unconditional payments to families to protect them against any financial loss from participating in the experiment; insurance against untoward events for participants; and clear wording of written materials and oral presentations in explaining the benefits and obligations of participation.

The doctrine of informed consent has traditionally been the ethical guideline; in the United States, however, institutional review boards (composed of scientific peers) have begun to supplement the experimenter's obligation to obtain informed consent. The review boards assess the risks to the participants and the benefits to society from the proposed experiment (or other data collection project). In many instances, governments will not grant financial support to a project if an institutional review board has given it a negative assessment.

Nonmeasurable Outcomes

Finally, an experiment will not be appropriate if important outcomes will not be measurable. Government agencies, especially, sometimes launch programs for symbolic purposes, or as proof of the government's concern about some issue or group of people (Glennan et al., 1978). Although these actions are sometimes

termed "experiments," they are often hastily put together, whereas the lead time required to field a scientifically sound experiment makes it unlikely that such actions will produce useful knowledge. For that reason, the principles expounded in this report are not meant to apply when the primary intent of the policymaker is symbolic.

II. ORGANIZATION AND MANAGEMENT OF AN EXPERIMENT

Writing about the management of a social experiment is much like writing about management in general. Although specific, substantive knowledge is useful, classical management skills are nonetheless required to operate a social experiment successfully. The experiment manager can be likened to the executive vice president of a private enterprise, who must possess a full range of planning, operations, financial management, personnel management, public relations, and communication skills.

Others who have written about social experimentation have stressed the importance of management to an experiment's success (Riecken et al., 1974). We heartily support this point. However, we look dubiously on the mystique that seems to have grown up around the management role in the literature. The literature tends to dramatize the innovative nature of social experimentation at the expense of the many mundane but vital tasks of operating an experiment. In our judgment, the management skills required for social experiments are not unique. Although special expertise is helpful, application of traditional organizational and managerial principles is essential.

The danger in our starting out with such an assertion is that it may mislead the reader into the belief that we are embarking on an "inspirational" treatise, in which we try to put new life in hoary management maxims, such as "plan ahead." Given our perspective, this is to a certain extent true—and inevitable—but there is much more to it than that.

The following sections elaborate on aspects of the organization and management of social experiments that we have found important.

ORGANIZATIONAL DESIGN

Because social experiments typically involve several institutions and extend over several years, the experimenter must explicitly consider organizational design at the very outset. The apportionment of responsibility and authority makes a difference; it is one of the experimenter's most important jobs, and it continues as the experiment evolves.

Although many organizational arrangements can be suitable for a social experiment, an important tenet is that a single organization should be vested with total responsibility and authority for completing the work—from experimental design through program operation and research. The generally large scale of social experiments, and the broad range of skills required, frequently lead to the formation of consortia and partnerships. These can be viable organizational forms, but the high degree of coordination that an experiment calls for suggests that ultimate authority be as unambiguous as possible. From a management perspective, none of the difficulties in assembling the requisite mixture of talent and job assignments should be allowed to weaken the concept of a sole source of accountability.

Even though one organization should have ultimate responsibility, a team approach to implementing an experiment is an absolute necessity. Riecken et al. (1974) have defined six roles that appear in the social experimentation process: initiator; sponsor; designer-researcher; treatment administrator; program developer; and audience-user. Field and Orr (in Boruch and Riecken, 1975) have pointed out that the initiator and sponsor are often one and the same, and that the roles of designer-researcher, treatment administrator, and program developer may be executed by people from the same organization. The primary reason for defining specific roles, however, is to stress the variation in perspective, motivation, and skills required in an experiment. It is unlikely that the staff of any one institution will have all the necessary skills. Even if it does, the task of coordinating these skills is far from trivial.

THE TEAM APPROACH

The team approach is designed to overcome the most frequent source of dissonance cited by those who have analyzed social experiments: a conflict between the *Weltanschauung* of the designer-researchers and of those who must carry out the day-to-day activities. This has been characterized as the conflict between the "research" team and the "action" team (Riecken et al., 1974).

The team approach is easy enough to describe, but hard to realize in practice. First, representatives of each component of activity must participate fully and as early as possible in the experimentation process. In particular, the "action" perspective should be represented in the design phase because the choice of experimental treatments must include assessments of their operational feasibility. (The data processing group should also be included in this phase.) Treatments not considered by the action team have a lower probability of success, not only because of objective difficulties, but also because the action team has less stake in their success.

For example, a treatment that routinely requires substantial judgment by an action team member may be so difficult to administer consistently that the team member loses faith in the experiment. This may happen because the team member is aware of the inconsistency and cannot conceive how the information gathered can be useful to the researcher. A probable consequence of such a loss in faith is even less consistent administration than might otherwise have been the case. Eventually it may result in an inferior treatment for the researcher to analyze. In this case, participation in the design by the action team may lead to a redesigned treatment with less action team judgment required. Or, equally important, the action team may be informed clearly that the treatment is supposed to include administrative judgment and that whatever inconsistency is incurred by well-intentioned administration will be measured and analyzed. In either case the participation in the design process by the action team should improve the experiment.

Second, the experiment's manager(s) must cultivate and perpetuate mutual respect for each perspective within and across participating organizations. Unproductive conflict between the two perspectives is a constant hazard. As Riecken et al. (1974) have stated, "Nothing is more crucial for the successful management and execution of the social experiment than good cooperation between action and research teams." Strong disagreements are bound to arise, however, and little can be

done to suppress them—and rarely, in our judgment, should they be. Rather, they should be turned into productive exercises that in the end contribute to the success of the experiment. Close, daily communication between the head of the research team and the head of the action team is essential.

Third, routine—as opposed to ad hoc—mechanisms must be designed for resolving differences. Such mechanisms should enable ready contact between action and research team members; this can promote a useful broadening of the perspectives of both parties that will be especially helpful when difficult differences in approach must be resolved. Routine interaction also makes for consistent decisions and actions. Reliance on ad hoc resolutions risks too much "hardening" of positions because of one group's unfamiliarity with the objectives and constraints facing the other group. Conflict and tensions are inevitable, and a strategy that waits for them to become intolerable to either side risks more than does a strategy that tries to resolve them as early as possible.

For example, research designs frequently call for repetitive measurement. At times these checks may seem demeaning or burdensome to the action team member. Having to call back people recently interviewed to ask their birthdate when the action team member is positive the initial interviews were correct can cause resentment of the team member and the people interviewed—especially if the reasons for the procedure are not clearly understood. Similarly, requests to the research team from the action team for seemingly simple advice on how to respond to irate experiment participants can cause the research team to lose faith in the ability of the action team and to make unnecessary or unproductive changes in design or analysis to attempt to compensate for perceived deficiencies in conducting field operations. In either case, routine mechanisms for interaction and conflict resolution can forestall resentment or loss of faith by either team that might jeopardize experimental goals.

Fourth, although sustained interaction among groups is essential, team members and groups must be given as much latitude as possible to carry out their designated tasks without interference. Considerable division of labor is necessary in a social experiment. Some of this division can and will take place naturally, but some must be managed—even forced. Members or groups within teams ultimately derive their value from the special talents they contribute to team performances. Too much interference by one team in the activities of the other is likely in the long run to be detrimental to the exercise of that talent—and thus to the experiment.

For example, research team instructions to the action team on how to order postage stamps, install telephones, or purchase desks do little to generate the respect of the action team. The action team is bound to consider itself capable of handling such routine administrative tasks. Similarly, efforts by the action team to alter data collection activities because it thinks the data so collected will be more suitable for analysis frustrates the research team and possibly the goals of the experiment. These examples may seem trivial, but our experience suggests that such actions are not atypical in social experiments.

The research team will not be as successful as it might if research possibilities seem constantly to be constrained by dictates of the action team. Similarly, the action team cannot carry out its tasks if guidance and direction from the research team appear unduly constraining. An exhortation to both sides to exercise tolerance may be all that is needed to smooth over difficulties. When push comes to

shove, however, this advice breaks down. The research team has ultimate responsibility for the experiment, and cannot relinquish that responsibility.

THE KEY FOCUS: PRESERVING THE INTEGRITY OF EXPERIMENTAL TREATMENTS

Of all the managerial challenges, perhaps the most compelling is the need to preserve the integrity of the experimental treatments in the face of both operational and analytic difficulties and opportunities. We have tried to apply a fundamental test whenever that challenge arises: To what extent do alternative courses of action strengthen or threaten the integrity of the experimental treatments? This question may seem mundane out of context; perhaps only those who have experienced the challenge of managing a social experiment will sympathize with the need to make this point explicitly.

In the absence of other guidelines, many operations decisions will be made to suit administrative convenience. However, social experimentation also requires considering the decision's effect on participant behavior and the resulting ability of the analyst to make inferences. Thus, the action team must think about research implications as well as the convenient or expedient method to accomplish a task. We do not deny the utility of administrative simplicity, but research implications are also important. Taking account of research implications is not easy, and not always successful; but the rewards for doing so are high. (Perhaps more accurately, the penalties for not doing so are high.) Training, well-thought-out procedures, and an organizational structure that facilitates sustained interaction between the action and research groups throughout the operation of the experiment will foster the goal of taking account of research considerations. Such a simple matter as placing the research and action teams in close physical proximity has proved very helpful. At a more general level, frequent use of the question "How will my actions affect participant behavior?" has been a useful guideline for action-team decisionmaking.

Such a guideline is helpful; unrealistic expectations are not, such as expecting the action team to be trained in statistical inference or to have more than a moderate appreciation of the technical research skills that will be used to analyze the experimental data.

One great danger is that the research team may never know what field decisions were made without considering the possible impact on analysis. Most social experiments involve field operations so extensive that written procedures cannot cover every situation; virtually all experiments have crises that appear to require immediate action. Thus, it is unrealistic to expect that the research team can participate in all the day-to-day decisionmaking of the field staff. Consequently, mechanisms must be designed to trigger communication between the teams when important issues arise.

How can the action team distinguish innocuous situations from those it should take to the research team? The action team should ask whether its actions may affect participant behavior in a fashion not already known to be accounted for. If the question is raised, the proper disposition will usually be clear. The moderately simple instruction to raise questions based on this guideline should help assure that

interteam conflict takes place only on significant issues and that the research team does not squander its time on routine matters.

Inevitably, the need for some crisis management and quick judgment will arise, and the actions taken are highly likely to affect other matters or actions taken later on. When time and resources are available to collect information for decisionmaking, these effects may well be foreseen; often, however, neither sufficient time nor resources are available for as full an investigation of a matter as one might desire. Even then, the effects may remain highly uncertain. Furthermore, experiment participants frequently have little tolerance for delays in decisions. Consequently, both the action and research teams can feel pressured to act quickly.

That happened to us, for example, during the Health Insurance Study when we got a telephone call in the late afternoon informing us that a participant, from whom we had collected 75 percent of the expected experimental data at a cost of more than $10,000, planned on withdrawing "tomorrow" unless we resolved his immediate family budget crisis. And, as might be expected, his case was a complex and "different" one that our standard guidelines did not cover. When faced with such decisions, our course of action is to gather as much information as possible in the limited time available and then speculate about the effect of alternative actions on treatment integrity. This approach can sometimes lead to unorthodox (even sloppy) administrative practice, and frequently complicates analysis (for example, preserving the integrity of the treatment may require gathering data outside of the established system and identifying the participant for special treatment by the analyst). Nevertheless, failure to remember the fundamental purpose of the experiment in the "heat of battle" can be extremely costly.

ADAPTABILITY: A NECESSARY INGREDIENT

The transformation of any experimental design into a field operation is likely to generate unanticipated consequences. Both the research design and operations plan must be adaptable. Protocols that are too strict are doomed to failure as the action team struggles to force human behavior into the prescribed treatment. Operations plans carried out only according to administratively precise procedures can irreversibly compromise the possibility of valid treatment comparisons before the first set of data reaches the research team, if those procedures later prove inappropriate. We reemphasize that research and action teams operate in different worlds. We have already stressed the importance of bridging the gap between these worlds. Adaptable experimental designs and field operations make this easier.

Expecting the unexpected, the experimenter must know when to modify design or operating procedures. Thus, defining good performance measures is important. How does the action team know when changes are necessary? Standard administrative performance measures are of little help. The experiment manager seeks consistent treatment of participants, uniform application of data-collection procedures, careful organization and delivery of services according to specifications, and excellent early warning systems of potential flaws in the design. Unfortunately, conventional efficiency measures, such as cost per unit output, number of units processed per month, or number of complaints per unit of output do little for the experiment manager.

In lieu of such indicators, the experimenter must design measures to judge the

effect of operations on participant behavior and the integrity of experimental treatments; the manager will then have the opportunity to make "mid-course" corrections. If the performance measures are ill defined or badly reported, the manager simply cannot know when to intervene. Performance measures are also needed to monitor the effect of any intervention. The manager must know if adjusting one part of the system is wreaking havoc with other parts. Such knowledge requires performance measures that allow sufficient oversight and control to confirm that global experimental objectives are being met.

Creating timely feedback and establishing a basis for understanding participant behavior can help in the design of such measures. Most data collection systems used in social experimentation result in machine-readable data that are batch-processed after considerable preparation and editing. This inevitably takes time, often creating delays of many months between initial data collection and subsequent analysis by the research team. In the intervening period, the manager needs mechanisms to get a quick reading on the success of data collection activities. Elaborate measures that reveal only after the fact how well the experiment was managed can aid the analysis, but they do not help the manager to intervene at the right times. Section III below discusses several analytical methods used in the Health Insurance Study to test our success in being able to make valid inferences from the experimental data. However, we also have used devices such as weekly progress reports, periodic system audits, site observation, and analysis of hand-tallied data to provide more immediate feedback about the progress of experimental operations.

We also have tried, when possible, to anticipate participant behavior and to be in a position to react when participant behavior appeared to deviate from our expectations. A simple case in point, when we failed initially but took corrective action, occurred when we mailed two complex self-administered questionnaires at the same time to one of our sites, but mailed them at different times to the other sites. Response rates were lower and data retrieval costs higher at the site to which we mailed the questionnaires together. We corrected the situation by separating the mailing periods for subsequent questionnaires. Since then, we have considered much more carefully how the burden on the respondent will affect the success of our data-gathering activities—even though we were aware of that factor from the beginning of our project. The point is that such signals should not be missed; continuous monitoring of the effect of operations on respondent behavior helps ensure that they will not be.

This example also illustrates the importance of an adaptable design and field operation. The questionnaires were mailed together in the first place because the design called for administration of one (an income report) in the spring, when income-tax forms are due in the United States, and the other (a health questionnaire) at a yearly interval after enrollment—which also happened to be in the spring at this site. Being able to adjust the interval because of the robustness of the design, and having an action team that could rearrange its activities without trauma, made this (rather minor) mid-course correction possible. The alternative without this flexibility may well have been to accept a lower quantity and quality of data.

Another specialized technique for creating useful intelligence is the use of extensive quality control systems. Because quality control approaches are common

to survey research and data processing, two key activities in any social experiment, we do not review their development and use here. Rather, we highlight quality control as a special case of designing performance measures to focus attention on data quality as well as administration.

Quality control plays two important roles in a social experiment. Because of the long period between data collection and analysis, field quality control operations must carry the principal burden of maintaining data quality. Most decisions regarding data quality will be effectively made at the time of collection—with little prospect for reversibility. Therefore, the research team should participate fully in the design of data quality control checks and monitoring of results.

Establishing a quality control system offers an opportunity for productive cooperation between the action and research teams. It is an especially difficult challenge for the researcher, who will be required to make judgments without having the opportunity to analyze data in any sophisticated manner, and who may be forced to make decisions that run some risk of compromising analytic possibilities. But intensive participation by the researcher at the data collection stage of an experiment can preserve analytic options that may otherwise be unintentionally foreclosed. In addition, the greater the participation of the researcher, the better the chances for appropriately adapting to changing or unforeseen circumstances.

Quality control procedures also act as important communication devices between the research and action teams. They permit the researcher to communicate intended outcomes to the action team in a concrete, useful manner. Quality control instructions signal to the action team what is important in the field operation and they help the action team react to field circumstances.

Certain measurement problems are common to many social experiments. Many economic experiments, for example, must deal with the problem of how to measure income. Considerable folklore and some research suggest ways to measure income, but not all the difficulties have been resolved. Lectures and elaborate statements of principle are rarely as successful as detailed procedures in communicating to the action team how to define income and when to seek policy advice from the research team.

PLANNING AND MANAGING THE INFORMATION FLOW

Experimental operations require day-to-day management of a large amount of information, something to which many researchers may be unaccustomed. Much of the information requires action. From a management perspective, this fact of life is not particularly troublesome; well-trained managers are expected to design and operate information systems without difficulty. However, the information flow does need to be planned and managed; if it is ignored for very long, it can get out of control and chaos may result.

The information flow, when properly structured, can serve to create a written record of what has already transpired in the experiment. Especially if administration of the treatments is complex and the experiment runs for several years, it is difficult if not impossible to retain needed information for decisionmaking in personal memories. The written record can help counter both threats to consistent application of the experimental program—which are legion—and changing per-

spectives that accompany the maturation of the experiment. We caution, however, that seemingly simple tasks, such as careful file organization, rigorous documentation, and inclusion of all relevant parties in decisionmaking, are difficult to maintain in practice.

A well-structured, written record helps in another way. Many decisions cannot be reduced to small, discrete issues susceptible to unilateral consideration. As a result, several parties will participate, and so more thorough and detailed information will be required to support the decisionmaking process than would be the case if decisions were made unilaterally.

The management information system not only has to provide timely, relevant information (as usual), but it also should simplify accounting for the effects of change in one part of the system on other parts. In communicating the consequences of changes in design or procedures to various teams, large-scale experiments will be forced to adopt some bureaucratic characteristics. Although we would not wish the well-known bureaucratic pathologies on an experiment, neither do we believe that most experiments can be run in the collegial style more familiar to academic researchers. The need to structure information flow, accept considerable division of labor, and purposefully design communication and decisionmaking procedures to account for the variety of activities and actors involved must be recognized at the outset.

SKILL REQUIREMENTS

We noted at the beginning of this section the wide set of skills that will be drawn upon in the course of a social experiment. Although these skills are not mysterious, we list them briefly here as a reminder that they should be required for experimental designers and managers.

Financial Management Skills

Financial skills include budgeting, cost-estimation, accounting, cash-flow management, and financial auditing. Naturally, a social experiment should have a budget. However, especially if the research or the treatments involve considerable uncertainty, "the" budget is likely to consist of many budgets, adjusted over time to reflect changing costs, the operating environment, and research progress.

One key difficulty facing the experiment manager is how to trade off research possibilities in the face of the uncertainty of the cost of field operations. Pessimistic cost estimates may suggest immediate budget constraints that do not materialize over the long run, leading the manager to cut back activities prematurely; or actions based on optimistic cost estimates may force later compromises to the research design that a more conservative approach could have avoided. Two implications follow: (1) Without careful management, cost estimates and budget realities can end up dictating the course of the experiment to a much greater degree than might be imagined; and (2) because the manager's tradeoffs between cost and research possibilities may have more irreversible consequences than do many other decisions, a high premium attaches to sound cost information. The timing of

adjustments to research plans, and consequently to field operations, can profoundly affect the richness of experimental outcomes.

The most important building blocks to a budget are good cost-estimating procedures. Cost uncertainty is likely to afflict both the administration of the experimental treatments and whatever survey activities are undertaken—to say nothing of the cost of the research. Also, in lengthy longitudinal studies, data processing accounts for a large portion of total expenditures. Managing an experiment therefore involves scores of cost estimates for various components of the experiment, and periodic projections of the total cost of the experiment. Revised budgets will have to be issued frequently to reflect the results of these analyses—perhaps as often as quarterly instead of the more common annual budget exercise.

Accounting and auditing skills are also important. Because experiments often are funded with public monies, there is an obligation to design financial systems that are easily audited and that organize expenditures into standard systems of accounts. The peculiarities of government contracting and accounting in the United States mean that grants or contracts may not be audited and "closed out" until several years after the expenditures have been made. Actions taken in the "heat of the battle" by the experiment manager should be supported routinely by sound accounting practices so that issues of liability never have to cause concern later.

One perhaps unexpected skill that some social experimenters have found necessary is cash-flow management. When large sums of money are involved, the timing of transfer of funds from one agency or unit to another can be important to the experiment manager. Managing that flow wisely can effectively increase the budget of the experiment.

Since these financial management techniques are commonplace to executives in their everyday dealings, there is no reason for a mystique to surround their use in social experiments. But familiar as they are, they represent a set of skills that must be incorporated into the team conducting a social experiment.

Coordination Skills

Most social experiments call for skills in planning, scheduling, and project management. In a general sense, these skills amount to providing leadership and coordination. In a more specific sense, they are central to the day-to-day operations of an experiment, and require more than exhortation. Not only do schedules and project plans affect costs, they inevitably set the pace for the analytic effort.

Long-range planning should be a natural part of the experimental design process. But just as a social experiment is likely to have many budgets, so is it likely to have many plans. The primary skill required is to integrate each component of experiment operations into the larger design of the experiment. This calls for detailed projections of resource needs, environmental constraints, task definitions, task completion times, and interrelationships. Although it is not likely that computerized networks and critical path algorithms such as those provided by PERT and CPM will prove cost-effective in most social experiments (the cost of input and update may well exceed the cost of more conventional methods), the basic principles underlying the use of these techniques will have to be applied in one form or another.

Scheduling activities will prove to be a nontrivial task, and at the very least Gannt charts and similar scheduling tools will be necessary. Many organizational units may require detailed schedules of their own, with relevant unit outputs or needed inputs forming the basis for a more global project schedule.

The particular technique used is less important than careful planning and scheduling; otherwise substantial delays and inefficiencies are likely. And, of course, any plans and schedules require monitoring and updating. Obviously, failure to do so penalizes the experimenter most severely when operations have uncertain outcomes or when the design of future activities depends on results from preceding activities.

Other Executive Skills

Brief mention should be made of a series of other skills that may be needed at one time or another during a social experiment. The need is likely to arise for public relations skills, personnel management skills, legal and contract negotiation skills, data processing skills, and, as usual, good written and oral communication skills.

People and their behavior are the subject of any social experiment. In reporting its results to a wide audience, the experimenter will want to explain clearly the purpose of the experiment, the details of treatments, the obligations of participants and cooperators, and a myriad of other details. Some programs will use multimedia campaigns and various communication techniques to recruit participants. (The Housing Allowance Supply Experiment conducted extensive "outreach" campaigns for that purpose.) In other cases, cooperation of various actors in the environment may be essential. (The Health Insurance Study could succeed only if physicians and other health care providers agreed to treat its participants.) In any case, experimenters have found it necessary to design, take part in, or at the minimum approve campaigns to "sell" their program. Researchers do not necessarily have the skill or the time to mount successful public relations campaigns.

Relations with the sponsor are also important for the experimenter, who will probably have frequent interaction with the sponsor. Most experiments deal with questions of public policy, and the researcher may find it hard to maintain the longer-term, more reflective view of the scientist if the sponsor is pressing for shorter-term results that are directly relevant to policy. During the course of the study, the experimenter is likely to need communication skills other than those that suffice for writing journal articles or final reports.

The number and diversity of personnel in some experiments will require more time and attention to personnel management than the researcher may be accustomed to. Also, a good deal of personnel turnover may occur in both the action team and the research team over the life span of prolonged experiments. Although much of the personnel function may be delegated to others, the experiment manager is likely to spend considerable time seeing that staff members have comfortable working conditions, appropriate compensation, and challenging responsibilities. Because many people will outgrow the jobs they held at the beginning of a long experiment, the manager will find it necessary to provide for their growth and development and find replacements for them when they move up or leave. Again, these functions are commonplace, but they may take much more time and cause more worry than a researcher may expect.

Legal skills will be used throughout the course of most social experiments. An analogy is the business enterprise that has counsel on a retainer, and uses services when necessary. Since legal advice usually is more valuable before the trouble or problem develops, rather than after, the experimenter should be alert to situations in which legal advice will be helpful and learn how to use it.

Many field operations involve routine legal processes such as obtaining business licenses, filing reports with regulatory or other administrative agencies, and checking for compliance with local laws. More substantial legal help may be needed in negotiating contracts or in establishing and maintaining data-safeguarding procedures.

It is not uncommon to subcontract work in a social experiment. Legal aid may be needed for writing and negotiating contracts, assuring conformance with government or other sponsor procurement regulations, and resolving disputes among parties. At least in the United States, an attorney is likely to be a standard member of the experiment team.

Perhaps most important, legal help may be needed in establishing procedures to protect the confidentiality of data provided to the experiment. Social science research in the United States has not reached the point where statutes, regulations, case law, or common practice provide an unambiguous guide for action. It is the goal of many experimenters to protect from third parties both the identity of participants and any personally identifiable data that they provide. In addition to being a matter of ethics, such protection may reduce refusal rates and improve the quality of data that the participants provide. Except in certain narrowly defined circumstances, however, full protection is not available in the United States. Information provided to the experiment can be subpoenaed, and may have to be made available to others. Disclosure can be minimized, but establishment of proper procedures, including the wording of any promises of confidentiality, will require legal help.

Data Processing Capability

Data processing capability is crucial to the success of a social experiment. Data processing design and management requires: (1) the organization of a data base in which all data collected by the experiment are stored; (2) creation of a sample-maintenance system to keep track of the status of individuals or institutions participating in or cooperating with the experiment; (3) design of a system to match the status (e.g., married, divorced) of persons, families, or institutions to the data collected from them; (4) design of an efficient method of extracting data and status information for many different combinations of variables of interest to the researcher; and (5) provision for dissemination of the data base to other researchers.

Typically, two generic types of data are collected in an experiment: survey data and program (treatment administration) data. Survey data include data from self-administered questionnaires or personal interviews conducted at one point in time, or at intervals, to gather sociodemographic or attitude information. Program data include information collected in the process of administering the experimental treatment (medical care expenditure information, income support payments, etc.).

Program data may be reported sporadically at the participant's initiative (for example, when a participant files a medical insurance claim), or may be collected routinely for program monitoring (for example, monthly income reporting as part of the calculation of monthly income maintenance benefits). Survey data are usually collected at the initiative of the experimenter. Although the data collection forms and procedures may be quite varied, at some point the data will need to be archived in a consistent, easy-to-access, machine-readable form. This is a significant challenge that requires skills in data base design and management.

But the task does not end with the organization of a data base. Perhaps the greatest challenge for the data processing system is to design a sample-maintenance system for keeping track of the status of each participant. At any one point in time a number of variables will define a person's status for the analyst (age, sex, employment status, marital status, location, etc.). During the course of an experiment, people's status will change—rather more frequently than might be expected. For example, they may change their address or employment, become separated or divorced, acquire children, leave the experiment voluntarily, or die.

The design of this system is enormously important to the experiment. The following example illustrates but one aspect of the problem. Suppose a three-person family is receiving health insurance benefits as an experimental treatment. The benefits are related to family income, and are paid when a member of the family files a health insurance claim. During the course of a year, the family splits up and forms two new families, one with two members and one with one member. However, neither family informs the experiment of the split until sometime after it has taken place. In the meantime, the experiment has paid some health insurance claims as if the original family were intact. Once the experiment learns of the change in family composition, how does it identify data collected between the time of the change and when it learned of the change? Alternative associations of family status with the program data will present different pictures to the researcher. The ideal system will provide information so the researcher can define family status in the manner most appropriate to the specific analysis at hand. However, the number of changes that can occur in the field make this ideal sample-maintenance system quite complicated. Providing for a wide range of options is easier said than done and will impose additional costs.

Given a good data base design and sample-maintenance system, the problem remains of matching the program and survey data to relevant participant status variables. Generally, the program data collection system cannot gather a full range of sociodemographic information (status variables) every time some program information is collected (e.g., is the participant employed on a certain date?). Some of this information will have been gathered from survey data or from periodic checks of program data. Typically, except when doing cross-sectional analysis, the analyst will always face uncertainty in associating data gathered from one form at one point in time with data gathered from another form at another point in time. The better the system for matching data to status variables (i.e., for tracking status across time), the higher the quality of analysis.

Most experiments have analytical agendas requiring different combinations of data for different analytical tasks. Even given consistent, efficient storage of data and ready matching to status variables, researchers must be able to extract different combinations of data for different tasks easily and cheaply.

In the United States, one of the selling points of experimentation has been the creation of large, longitudinal data bases that other researchers can eventually use to validate or extend the experiment's findings or to conduct research on new topics. This value of experimentation is lost if later researchers find it prohibitively expensive or complex to extract data from the data base for their own uses.

III. TIPS FOR THE EXPERIMENTER

Create a pilot sample. Typically feeling pressed, at the beginning, to produce results as soon as possible, the experimenter may be tempted to proceed rapidly with implementation. Despite the natural urge to get on with the work, the experimenter should establish a pilot sample but assume that the data from that sample probably will not be useful for formal analysis. The pilot sample should precede the regular sample by several months.

When purchasing a fleet of new aircraft, it is wise to "fly before you buy"—to develop a prototype to determine cost and feasibility before committing to a full-scale production run. Similar considerations apply to social experimentation. In particular, a pilot sample can:

1. Establish the feasibility of enrolling participants. It even may be possible to estimate two or three points on a "supply curve" of participants; i.e., discover how the refusal rate varies as a function of payments or obligations required of participants. Such knowledge may prevent either excessive refusal rates or monies paid to participants in excess of what the participant would require to enroll. Techniques for enrolling participants may also be tested.
2. Establish the ability to resolve procedural and definitional difficulties with the treatment. In some experiments, for example, there may be apparent legal problems.[1] A pilot sample should clarify the situation. More generally, the group responsible for conducting the experiment will gain experience they have probably not had in administering the treatment (e.g., in the income maintenance experiment, the experimenters had no prior experience in determining the amount of payment due to a participating family). By first administering the treatment on a small scale, lessons may be learned that will permit more efficient operation when scale increases. And it may turn out that the experimenters are incapable of operating the experiment, in which case a new group must be found or the experiment abandoned.
3. Provide a pretest group for testing interviews and other data collection instruments. Rarely can one design a flawless data collection instrument on the first try. It is axiomatic in survey design to pretest virtually all instruments. A pilot sample provides a convenient group on which to pretest such instruments.

A pilot sample, then, should show if an experiment is feasible: It can identify operational difficulties and problems with data operations and data collection instruments, and it will prevent the waste of resources that would occur if the first several months of data from an experiment without a pilot sample proved impossible to analyze because of continuing adjustments to unforeseen problems.

[1]For example, when we began the Health Insurance Study we did not know whether the experiment would be subject to state insurance regulation, and if so, what the consequences would be.

The experimenter should not regard the pilot sample as a means to demonstrate the analytical worth of the experiment or even, necessarily, to provide estimates of variances that will be useful in experimental design. Rather, the pilot sample should be used to test operational feasibility. Although the principle of using the pilot sample as the first step in a sequential experimental design appears attractive, it will probably not succeed in practice. There is a good chance that the experimenter will wish to change the protocol (and maybe even the treatment) after beginning the pilot sample. For example, results from pretesting interviews on the pilot sample may dictate the revision of interview instruments. It is unlikely that one can keep the data from the pilot sample comparable to data that will be generated later. Thus, although the statistical design of the experiment may benefit from the pilot sample, most of the benefit should accrue to operational aspects.

Build into the design an ability to measure effects that are an artifact of the experiment (methods or Hawthorne effects). At the conclusion of an experiment, those conducting it may have to deal with the common criticism that the participants' behavior might have been different in a real program because experiments and programs differ, if only in the amount of data collection. Anticipating that criticism, the experimenter can often build into the experimental design, at relatively little expense, an ability to detect and measure effects that may be peculiar to the experiment (methods effects). The standard technique for doing so is to collect data on two comparable groups, one that receives the experimental treatment and one that does not. Sometimes it will be impossible to comprehensively assess methods effects in the experimental design, but this does not absolve the experimenter from attempting to quantify those effects that are amenable to measurement.

Below are several examples that arose in the Health Insurance Study, illustrating the possibility of measuring experimental or methods effects. These examples and others are discussed at greater length in Newhouse et al. (1979).

1. At first, the Health Insurance Study (HIS) proposed to measure the health status of all participants by means of a medical screening examination at the beginning and the end of the experiment. This would have yielded much more precise information than would a single examination at the end. But one objective of the experiment is to measure how medical care use varies as a function of the health insurance plan. It is easy to see that a screening examination could affect that use. For ethical and legal reasons, for example, HIS staff are obliged to report any medical problems or abnormal symptoms uncovered by the examination to the physician the participant designates. The physician may well wish to follow up such results. Any resulting treatment would not have occurred without the screening examination; moreover, follow-up may be more aggressive if the participant is in a treatment plan that requires less out-of-pocket payment for physician visits.

To measure the amount of follow-up, the HIS sample was split into two groups at the time of enrollment: those who were to take an examination and those who were not. (Reflecting the importance of the data to subsequent analysis of health status changes, about 60 percent of the participants were given an examination.) Everyone is asked to take an examination at the end of the study, when follow-up will not contaminate the data.

2. Because an experiment usually lasts for only a limited period of time, families may not behave as they would if the experiment were longer (Metcalf,

1973; Burtless and Greenberg, 1978). The experimenter presumably wishes to generalize to steady-state behavior and therefore must estimate the effect of experimental duration on behavior. Two techniques are useful for the purpose: (a) The length of participation can be varied within the experiment; for example, in both the HIS and the Seattle-Denver Income Maintenance Experiment, some families participate for three years, others for five years. (b) Data may be collected after the experiment is over; comparison of these data with data from the experimental period permits estimates of transitory behavior. Such behavior may include "crowding in" during the experiment (e.g., purchasing medical care while it is cheap) or postponing certain actions until after the experiment is over. Arrow (1975) provides a rigorous discussion of the use of postexperimental data to estimate transitory behavior.

3. The HIS pays families for participating. The issue arose whether such money payments would cause families to act differently from how they would in a national insurance plan that did not pay such monies. Again, an "experiment-within-the-experiment" was set up; the amount of money paid to families was deliberately varied to measure the effect of the payments upon behavior.

Each of the preceding three examples illustrates an aspect of the experimental design that could alter behavior in a manner that would not be replicated in a general program. The most favorable outcome one could hope for from these experiments-within-experiments would be to demonstrate that certain measurement devices changed behavior negligibly, at most. Even if such an outcome obtained (and certainly if it did not), one would like a test of how much, if at all, participants altered their behavior because of their awareness of being observed or studied.

Such a test can be hard to perform, for it implies obtaining information on the behavior of a group of people (a "control group") without enrolling them in the experiment. Sometimes this will be impossible to do, because one simply cannot collect data without enrolling individuals, thereby making them aware that their behavior is being studied. On other occasions it will be possible to create a "control-on-control" group that is similar to a control group except it is not enrolled and does not receive the data collection instruments that a control group would.

Sometimes a one-time interview with a control-on-control group will serve this purpose. In the case of the HIS, for example, one may wish to know if asking a person about his or her health status annually for five years makes the person more health-conscious and therefore more likely to visit a physician. To test this hypothesis, one could in principle survey people not enrolled (or rely on existing surveys of similar populations) to ascertain their physician-visit rates. For such a method to succeed, data on people in the experiment should be collected by a similar method (that is, a similar survey document); otherwise, differences between the experimental and nonexperimental samples could be attributable to a difference in data collection methods.

Sometimes it may be possible to obtain information on a control-on-control group from records. Again, an example comes from the HIS. Part of the enrolled sample is a random sample of a prepaid group practice's members;[2] these people are formally enrolled in the experiment and are given all interviews. Their use of

[2]A prepaid group practice, sometimes called a Health Maintenance Organization, agrees to provide its members all necessary medical care for a fixed monthly price.

services at the prepaid group practice can be compared with that of other practice enrollees who are not enrolled in the experiment, using the medical records of the prepaid group practice. A second example is provided by the electricity rate experiment. This experiment enrolled a few hundred families in a control group and compared their consumption with that of a comparable group who were not enrolled, using billing records from the electric utility. Any systematic differences between these control groups should be attributable to the experiment.

Structure the experiment to keep refusal and attrition at low levels, especially refusal and attrition correlated with treatment. The desirability of keeping refusal and attrition at low levels is obvious, the means to do so less obvious. One way is to design the experiment so that no one is financially worse off from participating. In addition, payment for interviews can compensate the participant for the time and trouble taken to provide data.

Any payments made to minimize refusal and attrition rates cannot be conditional upon behavior the experimenter seeks to measure. The payments must be made solely for participating in the experiment, taking interviews when requested, mailing back self-administered questionnaires, and so forth. For example, the HIS families are paid a Participation Incentive, a sufficient amount of money to ensure that the families cannot be worse off financially from participating. Such a payment is necessary because the family's own health insurance plan may in some cases be better than the experimental plan to which it is randomly assigned. Without such a payment, then, the family would be worse off in some situations and it would be in its interests to withdraw.

Payments to families can be structured so as to give them an incentive to complete the experiment. In the HIS, part of the Participation Incentive is withheld until the end of the experiment, whereupon cooperating families also receive an additional amount, the Completion Bonus. Otherwise, situations could arise in which it is in the family's best interest to withdraw. One cannot completely eliminate attrition, of course. Random attrition at a low level should not be worrisome; it will cause only a small loss in efficiency of estimation. Nonrandom attrition, however, even at a seemingly low level (e.g., five percent of the sample), can lead to nontrivial bias. Methods for analyzing censored samples (Heckman, 1976; Hausman and Wise, 1976, 1977; Nerlove, 1978; Griliches, Hall, and Hausman, 1978) can be used to test for the possibility of nonrandom attrition and refusal. Unfortunately, such methods require the analyst to assume what the statistical distribution of observations would have been without the attrition, and the analyst usually has little or no basis for choosing among alternative assumptions. One can minimize dependence on such methods by designing the experiment to keep attrition at minimal levels.

Do not attempt to eliminate bias. Although prudence dictates an experimental design that will keep refusal and attrition—and thus possible bias—at a low level, it is almost certainly not desirable to attempt to eliminate bias; the expense is likely to be high, and the money could probably be used to greater advantage elsewhere.

Bias can arise in an attempt to balance families across treatments (as an efficient design requires), that is, to ensure that the distribution of families within treatments is as similar as possible for each treatment (see below). To achieve that balance, it is necessary to collect and process information about candidate families, and return to them later with an offer to participate on a specified treatment.

Unresolved issues of sampling are raised if the family's composition changes (someone moves in with the family; the heads separate; the family moves out of the area). Morris, Newhouse, and Archibald (1979) discuss these problems; it suffices here to note that it is virtually impossible in practice to maintain an unbiased sample frame. To do so, for example, would require following families who move from the area. Although that may be practical once the families have been enrolled in the experiment (using self-administered interview instruments), it is expensive to do so prior to enrollment. The amount of bias, if any, introduced into the sample by not following movers out of the area prior to enrollment (and, say, substituting into the sample the family that has moved into the dwelling) is almost certain to be small in most applications; it therefore will not be a wise use of funds to eliminate it.

Usually balance the sample across the treatments. Balancing the sample means ensuring that the characteristics of participants assigned to each treatment are as similar as possible to those of participants assigned to other treatments. An algorithm for balancing the sample, the Finite Selection Model, is now available and can be used for this purpose (Morris, 1979). Its use can achieve substantial gains in precision over simple random allocation of subjects to treatments, and gains over simple blocked allocation schemes may also be possible.

The income maintenance experiments in the United States used the Conlisk-Watts model to allocate families to treatments (Conlisk-Watts, 1969; Conlisk, 1973). This model can be used to determine the optimal number of families to receive each treatment, and we recommend its use in this manner. In the income maintenance experiments, however, the model was used in another way that had the effect of unbalancing the allocation of the sample to treatments. The experiments' designers used the Conlisk-Watts model to exploit the dependence between costs of different types of families and the treatments to which they are assigned. High-income families are relatively less expensive if placed on generous income maintenance plans because they receive fewer benefits than low-income families from a generous plan, whereas neither type of family receives many benefits from a stingy plan. (A generous plan has a high guarantee and low tax rate.) As a result, high-income families were disproportionately allocated to generous plans. If the effect of the plan is independent of income (that is, there is no interaction between income and treatment), this disproportionate allocation will improve precision. But the assumption of no interaction may be incorrect, and the unbalanced allocation makes it difficult to test the assumption (Keeley and Robbins, 1978; Cogan, 1978). This problem can be avoided by balancing the sample.

Although we recommend use of the Finite Selection Model to assign families to treatments, the important point is to ensure a robust design. A balanced design will be robust; that is, it should prove suitable not only for analyses where the designer is fairly confident that there is no dependence between treatment and family characteristics, but also for unforeseen analyses where such an assumption may be less appropriate or may be questioned by others.

Do not strongly oversample a group whose membership is not well defined. The experimenter is often more interested in a program's effect on one population subgroup than on another. It is appropriate to oversample the subgroup of interest (the favored group) if it can be defined with little error (for example, people more than 65 years old). But if a favored group can be defined *a priori* only with substantial error (for example, those who will consistently have incomes less than $5,000

per year), disproportionate sampling can reduce the experimenter's precision of estimation, even for the favored group (Morris, Newhouse, and Archibald, 1979). We recommend that the experimenter ascertain the reliability of any measures used for disproportionate sampling before setting sampling fractions, and that sampling fractions not depart far from proportionality if the classification contains measurement error.

Go to some lengths to inform participants about the treatment. Informed consent should be an ethical precept for an experimenter. Apart from the ethical argument, however, the experimenter's self-interest generally dictates fully informing the participant about the details of an experimental treatment. First, full information at the time of enrollment will help minimize attrition and, in a well-designed experiment, refusal rates as well. Attrition should be low because participants will receive no unpleasant surprises after enrollment; refusals should also be few because proper experimental design and full information should make it clearly in the participant's interest to participate. Second, the experiment will almost certainly last for only a limited time (rarely more than a few years). If behavior under the experiment is to be generalized to behavior under some kind of permanent or open-ended program, it is important that the participants understand the experiment as fully as it will the later programs. To convey such understanding, the treatment should be explained to the participant as clearly and simply as possible.

Choose the number of sites and length of enrollment to minimize variance, given a fixed project budget. If there are no fixed costs to operating a site, the optimal sample will be completely dispersed, that is, not concentrated in a small number of geographic areas. Usually, however, there will be fixed costs for each site. For example, a sampling plan for the site may need to be created, and a field office may be necessary during experimental operations. If there are fixed costs for a site, there is a tradeoff between the number of sites (locations of participants) and the total number of participants. This tradeoff can be made to minimize the variance of variables to be estimated.

In unpublished work, Morris has shown that the optimal number of sites, K, equals $BV^{1/2}/F(+ V^{1/2})$, where B is the budget available, F is fixed costs per site (assumed invariant to the number of sites), and $V = (F/M)[t^2(1 - r + rL)/s^2d]$, where M is the marginal cost per family (assumed constant), t^2 and s^2 are between-site and within-site variance, respectively, r is the correlation coefficient among treatment means within a site, L is the proportion of "level" estimates desired as opposed to "treatment contrasts" (a level estimate is the mean for each treatment; a contrast is the difference between treatments), and d is $\Sigma p_i c_i / \Sigma p_i^2 c_i$, where p_i is the sample proportion on the i^{th} treatment and c_i is the relative cost of the i^{th} treatment. The parameter d is the effective number of treatments of equal interest (d equals the number of treatments if sample size and cost per treatment are equal for each treatment; otherwise it is less than the number of treatments).

The optimal number of sites, K, rises with the total budget available and falls as fixed costs per site, F, rise; K rises as between-site variance rises relative to within-site variance. A larger value of r means that treatment differences in one site can be used to predict differences in other sites, and therefore fewer sites are needed if one is interested in differences, but this correlation is irrelevant if one is interested only in levels (L = L).

Length of enrollment is typically a difficult variable to optimize *a priori.* In general, the greater the intertemporal correlation in the variable of interest, the greater the gain from a larger sample participating in the experiment for a shorter period of time. If there is zero intertemporal correlation (unlikely in practical cases), a small sample participating for a long period of time will ultimately provide the same information as a large sample participating for a short period of time. Because the large sample participating for a short period will provide data earlier, however, it should be preferred. But three potential problems must be noted. First, if one wishes a measure of the long-term effect of the treatment (as one usually will), participants must be given enough time to adjust their behavior in response to the experimental treatment. This may require operating the experiment for several years. Second, as explained above, it is frequently desirable to vary the period of participation in order to measure the effects on behavior of a limited time of enrollment. The proportion participating for a longer period should reflect the weight accorded to determining the experimental effect of length of participation, relative to the weight given to having results early and having results with less variance when there is intertemporal correlation. Third, for an equal number of person-years, costs will probably be lower if a larger number of families is enrolled for a shorter period of time. Although there are one-time enrollment costs for each family, management costs are incurred for lengthening the duration of the experiment. In general, the additional management costs of a longer experiment dominate the additional enrollment costs.

Do not attempt too much. Most experiments have multiple objectives and use interdisciplinary research teams. Data gathered to serve one discipline or type of analysis may not be suitable for another. Assuming that resources are scarce, tradeoffs are inevitable. Naturally, the needs of the sponsor, the skills and preferences of the analyst, and the feasibility of alternative courses of action affect the manager's resource allocation decisions. But one should guard against attempting to do too much with one experiment, thereby risking doing too little. We advise the experiment manager to concentrate principally upon data collection activities that support sound analysis of treatment comparisons. Only after that is assured should one invest in potentially interesting methodological studies or auxiliary use of the data.

Although such a caveat may seem banal, our experience suggests that the unique setting of most social experiments can induce those responsible for managing it to overextend themselves. And once activated, the cycle of mistakes can be unforgiving. An increased research agenda usually demands increased field operations. All of the social experiments we are familiar with have had operation staffs functioning under extreme time pressure and severe budget constraints. When care is not exercised with the research agenda, performance can deteriorate severely in field operations. Indeed, the field operations may not be able to recover from an overambitious research design, thus risking contamination of the primary treatment comparisons.

Experiment managers should be conscious of their risk-taking behavior. They are likely to feel a pull between using accepted, "safe" analytical approaches and more ambitious approaches that push the state of the art. Alas, analysis plans that truly push the state of the art will fail some of the time. To guard against overextension, but still engage in some research that pushes the state of the art, we have tried

to do two things. First, whenever possible we avoid research activities that threaten our ability to complete the primary tasks of treatment comparisons. Second, we have tried to apply the concept of "nested" research, whereby results from small pieces of work feed into larger and larger tasks in such a way that failure in one task need not threaten the ultimate success of the important, broader tasks.

A key aspect of this approach is planning "fallback positions." In the HIS, for example, we have collected data to calculate a price index for each of our sites, even though existing regional data on prices could be used as an approximation if such an approach turned out to be infeasible. Our analysis will be better with the local index, but the viability of the experiment is not threatened without it.

Do not be easily discouraged. We believe social experimentation, if properly used, can be an extraordinarily valuable tool. To be sure, it is time-consuming, sometimes frustrating, expensive by the usual standards of social science research, and risky—a mistake in the design or its application may vitiate the entire endeavor. But new knowledge is seldom easily achieved.

BIBLIOGRAPHY

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

Arrow, Kenneth J., "Two Notes on Inferring Long Run Behavior from Social Experiments," The Rand Corporation, P-5546, September 1975.

Bawden, D. Lee, and William S. Harrar, *Rural Income Maintenance Experiment: Final Report,* Institute for Research on Poverty, Madison, Wis., 1976.

Boruch, Robert F., "Bibliography: Illustrative Randomized Field Experiments for Program Planning and Evaluation," *Evaluation,* Vol. 2, 1974, pp. 83-87.

——, and Henry W. Riecken (eds.), *Experimental Testing of Public Policy: The Proceedings of the 1974 Social Science Research Council Conference on Social Experiments,* Westview Press, Inc., Boulder, Colo., 1975.

Boruch, Robert F., et al., "Randomized Experiments for Evaluating and Planning Local Programs: A Summary on Appropriateness and Feasibility," in Howard E. Freeman (ed.), *Policy Studies Review Annual,* Vol. 2, Sage Publications, Beverly Hills, Calif., 1978.

Burtless, Gary, and David Greenberg, "The Limited Duration of Income Maintenance Experiments and Its Implications for Estimating Labor Supply Effects of Transfer Programs," Office of Income Security Policy, OASPE, DHEW, Technical Analysis Paper No. 15, Washington, D.C., October 1978.

Cain, Glen, and Harold Watts, *Income Maintenance and Labor Supply,* Markham, Chicago, 1973.

Campbell, D. T., and J. C. Stanley, *Experimental and Quasi-experimental Design for Research,* Rand McNally, Chicago, 1966.

Cochran, W. G., and Gertrude M. Cox, *Experimental Designs,* 2d ed., John Wiley and Sons, New York, 1957.

Cochrane, A. L., *Effectiveness and Efficiency: Random Reflections on Health Services,* Nuffield Provincial Hospitals Trust, London, 1972.

Cogan, John F., *Negative Income Taxation and Labor Supply: New Evidence from the New Jersey-Pennsylvania Experiment,* The Rand Corporation, R-2155-HEW, 1978.

Conlisk, John, "Choice of Response Functional Form in Designing Subsidy Experiments," *Econometrica,* Vol. 41, No. 4, July 1973, pp. 643-656.

——, and Harold Watts, "A Model for Optimizing Experimental Designs for Estimating Response Surfaces," *Proceedings of the Social Statistics Section,* American Statistical Association, 1969, pp. 150-156.

Cook, Thomas D., and Donald T. Campbell, "The Design and Conduct of Quasi-Experiments and True Experiments in Field Settings," in *The Handbook of Industrial and Organizational Psychology,* Rand McNally, Chicago, 1976.

Cox, D. R., *Planning of Experiments,* John Wiley and Sons, New York, 1958.

Ellwood, Paul M., Jr., "The Health Care Alliance," in *Report of the National Commission on the Cost of Medical Care, 1976-1977,* Vol. 2, American Medical Association, Chicago, 1978.

Enthoven, Alain, "Consumer Choice Health Plan," *New England Journal of Medi-*

cine, Vol. 298, Nos. 12 and 13, March 23 and 30, 1978, pp. 650-658 and 705-720.

Federov, V. V., *Theory of Optimal Experiments,* Academic Press, New York, 1972.

Ferber, Robert, and Werner Z. Hirsch, "Social Experimentation and Economic Policy: A Survey," *Journal of Economic Literature,* Vol. 16, No. 4, December 1978, pp. 1379-1414.

Field, Charles G., and Larry L. Orr, "Organization for Social Experimentation," Chap. 4 in Robert F. Boruch and Henry W. Riecken (eds.), *Experimental Testing of Public Policy: The Proceedings of the 1974 Social Science Research Council Conference on Social Experiments,* Westview Press, Inc., Boulder, Colo., 1975.

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

Gilbert, John P., Richard J. Light, and Frederick Mosteller, "Assessing Social Innovations: An Empirical Base for Policy," in Carl A. Bennett and Arthur A. Lumsdaine (eds.), *Evaluation and Experiment: Some Critical Issues in Assessing Social Programs,* Academic Press, New York, 1975.

Glennan, T. K., Jr., et al., *The Role of Demonstrations in Federal R&D Policy,* The Rand Corporation, R-2288-OTA, May 1978.

Gramlich, Edward M., and Patricia P. Koshel, *Educational Performance Contracting: An Evaluation of an Experiment,* The Brookings Institution, Washington, D.C., 1975.

Griliches, Zvi, B. H. Hall, and Jerry A. Hausman, "Missing Data and Self-Selection in Large Panels," in Marc Nerlove (ed.), *The Econometrics of Panel Data, Annales de l'insee,* Paris, April-September 1978.

Guttentag, Marcia, "Evaluation of Social Intervention Programs," *Annals of the New York Academy of Sciences,* 1973, pp. 3-13.

——, "Evaluation and Society," *Personality and Social Psychology Bulletin,* Vol. 3, 1977, pp. 31-40.

Hamrin, Robert D., "OEO's Performance-Contracting Project: Evaluation Bias in a Social Experiment," *Public Policy,* Vol. 22, Fall 1974, pp. 467-488.

Hausman, Jerry A., and David A. Wise, "The Evaluation of Results from Truncated Samples: The New Jersey Income Maintenance Experiment," *Annals of Economic and Social Measurement,* Vol. 5, No. 4, Fall 1976, pp. 421-445.

——, "Social Experimentation, Truncated Distributions, and Efficient Estimation," *Econometrica,* Vol. 45, No. 4, May 1977, pp. 919-938.

Heckman, James F., "Sample Selection Bias as a Specification Error," *Econometrica,* Vol. 47, No. 1, January 1979, pp. 153-161.

John, J. A., and M. H. Quenouille, *Experiments: Design and Analysis,* 2d ed., Charles Griffen and Co. Ltd., London, 1978.

John, Peter W. M., *Statistical Design and Analysis of Experiments,* The Macmillan Company, New York, 1971.

Keeler, Emmett, Joseph P. Newhouse, and Charles E. Phelps, "Deductibles and the Demand for Medical Care Services: A Theory of a Consumer Facing a Variable Price Schedule Under Uncertainty," *Econometrica,* Vol. 45, No. 3, April 1977, pp. 641-655.

Keeley, Michael C., et al., "The Estimation of Labor Supply Models Using Experimental Data," *American Economic Review,* Vol. 68, No. 5, December 1978, pp. 873-887(a).

——, "The Labor Supply Effects and Costs of Alternative Negative Income Tax Programs," *Journal of Human Resources*, Vol. 13, No. 1, Winter 1978, pp. 3-36(b).

Keeley, Michael C., and Philip K. Robins, "The Design of Social Experiments: A Critique of the Conlisk-Watts Assignment Model," SRI International Center for the Study of Welfare Policy, Research Memorandum 57, Menlo Park, Calif., November 1978.

Kehrer, Kenneth C., *The Gary Income Maintenance Experiment—Summary of Initial Findings*, University of Indiana (Northwest), Gary, March 1977.

Kempthorne, Oscar, *The Design and Analysis of Experiments*, John Wiley and Sons, New York, 1952.

Kendall, Maurice G., and Alan Stuart, *The Advanced Theory of Statistics:* Vol. 3, *Design and Analysis and Time Series*, Hafner Publishing Company, New York, 1968.

Kershaw, David N., "A Negative Income Tax Experiment," *Scientific American*, Vol. 227, No. 4, October 1972, pp. 19-25.

——, and Jerilyn Fair, *The New Jersey Income Maintenance Experiment:* Vol. 1, *Operations, Surveys, and Administration*, Academic Press, New York, 1976.

Kessner, D. M., et al., *Infant Death: An Analysis by Maternal Risk and Health Care*, Institute of Medicine, National Academy of Sciences, Washington, D.C., 1973.

Manning, Willard G., Bridger M. Mitchell, and Jan P. Acton, *Design of the Los Angeles Peak Load Pricing Experiment for Electricity*, The Rand Corporation, R-1955-DWP, November 1976.

Metcalf, Charles E., "Making Inferences from Controlled Income Maintenance Experiments," *American Economic Review*, Vol. 63, No. 3, June 1973, pp. 478-483.

Morris, Carl, "A Finite Selection Model for Experimental Design of the Health Insurance Study," *Journal of Econometrics*, Vol. 11, No. 1, September 1979, pp. 43-61.

——, Joseph P. Newhouse, and Rae W. Archibald, "On the Theory and Practice of Obtaining Unbiased and Efficient Samples in Social Surveys and Experiments," in Vernon Smith (ed.), *Experimental Economics*, JPI Press, Westport, Conn., 1979.

National Commission on Technology, Automation, and Economic Progress, *Technology and the American People*, Government Printing Office, Washington, D.C., 1966.

Nerlove, Marc (ed.), *The Econometrics of Panel Data, Annales de l'insee*, 30-31, Paris, April-September 1978.

Newhouse, Joseph P., "A Design for a Health Insurance Experiment," *Inquiry*, Vol. 11, No. 1, March 1974, pp. 5-27.

Newhouse, Joseph P., et al., "Measurement Issues in the Second Generation of Social Experiments: The Health Insurance Study," *Journal of Econometrics*, Vol. 11, September 1979, pp. 117-129.

Newhouse, Joseph P., and Charles E. Phelps, "New Estimates of Price and Income Elasticities," in Richard N. Rosett (ed.), *The Role of Health Insurance in the Health Services Sector*, National Bureau of Economic Research, Universities-National Bureau Conference Series No. 27, New York, 1976.

Newhouse, Joseph P., and Vincent Taylor, "How Shall We Pay for Hospital Care?"

The Public Interest, No. 23, Spring 1971, pp. 78-92.

Orr, Larry L., "The Health Insurance Study: Experimentation and Health Financing Policy," *Inquiry,* Vol. 11, No. 1, March 1974, pp. 28-39.

——, Robinson G. Hollister, and Myron J. Lefkowitz (eds.), *Income Maintenance: Interdisciplinary Approaches to Research,* Markham, Chicago, 1971. (See esp. papers by Orr and Kershaw.)

Palmer, John L., and Joseph A. Pechman (eds.), *Welfare in Rural Areas: The North Carolina-Iowa Income Maintenance Experiment,* The Brookings Institution, Washington, D.C., 1978.

Pechman, Joseph A., and P. Michael Timpane (eds.), *Work Incentives and Income Guarantees: The New Jersey Negative Income Tax Experiment,* The Brookings Institution, Washington, D.C., 1975.

Phelps, Charles E., "Demand for Reimbursement Insurance," in Richard N. Rosett (ed.), *The Role of Health Insurance in the Health Services Sector,* National Bureau of Economic Research, Universities-National Bureau Conference Series No. 27, New York, 1976.

Plott, Charles R., "Experimental Methods in Political Economy: A Tool for Regulatory Research," unpublished paper, n.d.

The President's Commission on Income Maintenance Programs, *Poverty Amid Plenty: The American Paradox,* Government Printing Office, Washington, D.C., 1969.

Riecken, Henry W., "Principal Components of the Evaluation Process," *Professional Psychology,* November 1977, pp. 392-410.

Riecken, Henry W., et al., *Social Experimentation: A Method for Planning and Evaluating Social Intervention,* Academic Press, New York, 1974.

Riecken, Henry W., and Robert F. Boruch, "Social Experiments," *Annual Review of Sociology,* Vol. 4, 1978, pp. 511-532.

Rivlin, Alice M., "How Can Experiments Be More Useful?" *American Economic Review,* Vol. 64, No. 2, May 1974, pp. 346-354.

——, and P. Michael Timpane (eds.), *Ethical and Legal Issues of Social Experimentation,* The Brookings Institution, Washington, D.C., 1975.

——, *Planned Variation in Education: Should We Give Up or Try Harder?* The Brookings Institution, Washington, D.C., 1975.

Roos, Leslie L., Jr., Noralou P. Roos, and Barbara McKinley, "Implementing Randomization," *Policy Analysis,* Vol. 3, No. 4, Fall 1977, pp. 547-559.

Roos, Noralou P., "Contrasting Social Experimentation with Retrospective Evaluation: A Health Care Perspective," *Public Policy,* Vol. 23, No. 2, Spring 1975, pp. 241-257.

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

Scheffe, Henry, *The Analysis of Variance,* John Wiley and Sons, New York, 1959.

Srivastava, Jagdish N. (ed.), *A Survey of Statistical Design and Linear Models,* North-Holland Publishing Company, Amsterdam, 1975.

Suchman, E. A., *Evaluation Research,* Russell Sage Foundation, New York, 1967.

Watts, Harold W., and Albert Rees, *The New Jersey Income Maintenance Experiment,* Vol. II, *Labor Supply Responses,* Academic Press, New York, 1977.

——, *The New Jersey Income Maintenance Experiment,* Vol. III, *Expenditures,*

Health, and Social Behavior, and the Quality of the Evidence, Academic Press, New York, 1977.

Weiss, Carol H., *Evaluation Research,* Prentice-Hall, Englewood Cliffs, New Jersey, 1972.

Wilson, John O., "Social Experimentation and Public-Policy Analysis," *Public Policy,* Vol. 22, Winter 1974, pp. 15-37.