An Overview of Social Experimentation and the Digest

This book testifies to the breadth and depth of a distinctive modern form of social research activity—the randomized social experiment. Such experiments have been conducted since the late 1960s to evaluate proposed changes in program or policy. Some experiments have been large and highly publicized—among them the Seattle-Denver Income Maintenance Experiment, the RAND Health Insurance Study, and the more recent experimental evaluations of state-run welfare-to-work programs and of programs run under the Jobs Training Partnership Act of 1982; others have been small and obscure. Some have “pilot tested” major innovations in social policy; others have been used to assess incremental changes in existing programs. A few have provided the basis for evaluating the overall efficacy of major existing programs. Most have been used to evaluate policies targeted at disadvantaged population groups.

This Digest contains brief summaries of 240 known completed social experiments. In addition, for purposes of contrast, we also provide summaries of 11 completed quasi experiments. Each summary, typically two or three pages long and presented in a standardized format, outlines the cost and time frame of the demonstration, the treatments tested, outcomes of interest, sample sizes and target population, research components, major findings, important methodological limitations and design issues encountered, and other relevant topics. The experiments summarized are those for which findings were available by April 2003. In addition, very brief outlines of 21 experiments and one quasi experiment still in progress at that time are also provided.

This introduction provides background information for readers regarding social experiments. Rather than attempting to comprehensively discuss social experimentation—a topic that would require an entire book of its own—we touch upon a number of areas pertinent to interpreting the Digest’s summaries. We begin by providing our definition of social experiments. This is important because we
used it in determining the evaluations to include in this volume. This is followed by discussions of the concepts of internal and external validity of experimental findings and the categories into which experiments tend to fall. We then briefly describe quasi experiments, noting their strengths and weaknesses. Next, we examine the reasons for conducting social experiments, provide an overview of ethical issues, and describe nonexperimental methodologies that have been proposed as substitutes. Some common threats to the external validity of social experiments are then reviewed, as well as “optional” features often found in experiments. A discussion of the uses of social experiments in the policy process follows. We then present a brief history of social experiments. The final section of this introduction explains the uses and organization of this volume, so that readers can make optimal use of the summaries.

WHAT IS A SOCIAL EXPERIMENT?

The summaries in this Digest focus on field studies of social programs in which individuals, households, or (in rare instances) firms or organizations were randomly assigned to two or more alternative treatments. The primary research objective of the experiments was to measure the effects, which are often called “impacts,” of the alternative treatments on market behavior (such as the receipt of earnings) and corresponding government fiscal outcomes (such as the receipt of transfer benefits). Thus, a social experiment has at least the following features:

*Random assignment*: Creation of at least two groups of human subjects who differ from one another by chance alone.

*Policy intervention*: A set of actions ensuring that different incentives, opportunities, or constraints confront the members of each of the randomly assigned groups in their daily lives.

*Follow-up data collection*: Measurement of market and fiscal outcomes for members of each group.

*Evaluation*: Application of statistical inference and informed professional judgment about the degree to which the policy interventions have caused differences in outcomes between the groups.

Random assignment involves neither choice nor discretion. Whereas human subjects may or may not have the right to choose to participate in the experiment, they do not have the right to decide which group within the experiment they will join. Similarly, those administering the policy intervention may restrict eligibility for participation in the experiment, but once a person is admitted, program staff cannot determine the group in which that subject is enrolled, except by using randomization.

Social experiments test policy interventions: They are attempts to influence the endowments of, and the incentives and disincentives facing, human subjects. Thus, in most social experiments, one of the randomly assigned groups, the control group, represents the status quo and is only eligible for benefits and services under the existing policy regime. The remaining group or groups, the treatment group(s), are subjected to the policy innovation or innovations being tested. Comparisons of the control group with the treatment group(s) indicate the impacts of the tested innovations.

Social experiments are designed to determine whether (or how much) the policies being tested would affect the market behavior of individuals (e.g., their employment and earnings, consumption of food, energy, housing, and health care services; receipt of government benefits). Taken together, the second and third features of our definition exclude random-assignment experiments in medicine, psychology, economics, criminology, and education that are not designed to measure changes in subjects’ transactions in their daily environment in response to policy innovations. For example, we exclude randomized clinical trials of
prescription drugs intended to affect health status; we also exclude randomization of students to competing school curricula intended to improve scores on standardized tests.

Although outcome data may be collected by a variety of means, readers of this volume’s summaries will notice that over time social experimenters have relied increasingly on administrative data rather than surveys. Once collected, outcomes among the groups of randomly assigned individuals are compared. Outcome differences between the groups provide estimates of the impacts of the tested policy interventions. However, the data do not speak for themselves. Analysts must decide what data transformations are appropriate; what, if any, nonexperimental factors should be considered; what statistical techniques to use; and what results do and do not make sense.

Internal and External Validity

If implemented properly, the results of social experiments generally are internally valid, that is, they provide unbiased impact estimates for targeted people subject to different treatments at the particular time and place they were administered. However, evaluation, including that based on random assignment designs, requires considerable care and is often expensive. A number of the experiments summarized here were done “on the cheap” and hurriedly. Many of these experiments tended to be poorly designed and implemented, and as a result, some of them either produced findings that lacked internal validity or did not produce findings at all.

Even when carefully implemented, the extent to which social experiments possess external validity—applicability to other individuals, places, and times—may be controversial and is always problematic. For example, differences in timing imply that social attitudes, government institutions, the business cycle, the relative demand for unskilled and skilled labor, the rate of inflation, and other factors may vary from what they were when the experiment was conducted. Different locations may result in dissimilarities in age, sex, racial, or ethnic mixes; social attitudes; state and local government institutions; industrial structure; and many other factors.

Categories of Social Experiments

Several of this volume’s social experiments—notably the four income maintenance experiments, the Housing Allowance Demand Experiment, and the RAND Health Insurance Study—were designed to estimate “response surfaces”—that is, their designs allowed two or more continuous parameters of the program being tested to vary within wide ranges. For example, tax rates and guarantee levels in the income maintenance experiments varied greatly across treatment groups. Estimates of responses to program parameters can, at least in principle, be used to project the effects of any program that has the basic features of the one tested, even if the specific values of the program parameters differ.

In contrast to response-surface experiments, most social experiments permit only “black box” assessments of whether the tested intervention “works.” That is, they provide different randomly selected groups of individuals with different “packages” of services and incentives (for example, job training, child care assistance, and job search assistance) and then determine whether outcomes (for example, postprogram earnings) differ among the groups. Only limited information is typically provided on the degree to which these impacts can be attributed to specific components of the service packages or on the effects of changes in program design. Findings from the response-surface technique are in some ways more flexible for projecting the effects of future policy, but it may not be possible in practice to carry out intervention(s) in the form of variations in the values of two or more continuous variables.
A second important distinction is between experiments that are mandatory and those that are voluntary. An experiment is mandatory if the individual cannot enjoy certain benefits without participating. Many unemployment insurance and welfare experiments in this volume have been mandatory. The Minnesota Income Tax Compliance Experiment, in which certain taxpayers were randomly selected for a higher probability of audit, was mandatory in a broader sense, because the individuals had not applied for any specific benefits.

Individuals must in some way apply to enter a voluntary experiment. Experimental evaluations of training programs and electric-rate experiments have usually been voluntary. As demonstrated later, the fact that individuals chose to enter an experiment may complicate the evaluation of the outcomes.

Government agencies often consider themselves legally and ethically justified in conducting mandatory experiments if the alternative treatments are within the agency’s ordinary administrative discretion. For example, randomly selecting certain taxpayers for audits is a necessary function of a tax office. Voluntary experiments generally require some type of informed consent by the subject.

A third distinction between experiments concerns the sort of policy intervention they test. There are three possibilities: they can test new programs; they can test incremental changes to existing programs; and they can be used to evaluate existing programs, although they are relatively rarely used for this last purpose.

WHAT IS A QUASI EXPERIMENT?

In addition to summarizing randomized social experiments, this book also contains outlines of quasi-experimental demonstrations. Scholars and evaluation professionals have attached several conflicting meanings to the term quasi experiment; to keep the scope of this book manageable, we have limited the term to policy demonstrations in which potential sites are randomized to treatment or control group status, and individuals or households at that site are clustered together for analysis. An example illustrates.

Suppose that six sites are considered for a demonstration, and some 1,500 individuals will be subjects of the experiment at these sites. On some prior basis, the evaluators determine beforehand that site A is most like site B, site C most similar to site D, and site E is best matched with site F. By flipping a coin (or using a more sophisticated random process), the evaluators determine whether the innovative treatment will occur at site A and the control at site B or vice versa at site C versus site D; and so on. For concreteness, let the treatment sites be A, D, and F, and the control sites be B, C, and E. The intention is to compare the ADF outcomes with the BCE outcomes and, if they are significantly different, to attribute that difference to the policy intervention.

Significant theoretical and practical reasons may exist for choosing site randomization over randomizing individuals within a site. For example, an innovation may be intended to change "the culture of the welfare office" or "the culture of public housing projects"; for this cultural change to occur, the innovation must affect all similarly situated individuals who are members of the culture. Feedback and information processes, which may be crucial (and are treated at greater length later in this introduction), might only function if an innovation is adopted on a sitewide basis. The responses on "the other side of the market" from employers or landlords may also be important, and randomization within sites might attenuate these responses. For any number of reasons, it also may be simply impractical to administer different treatments within the same site.

The most serious problem with the site randomization concept is that there may be too few sites for effective randomization to occur. Although
the sites have been assigned to treatment and control status by chance, the people within the sites have not been randomized. People do not randomly choose whether to live at site A or at site B; they choose purposefully, and their choices create both observed and unobserved differences between sites subject to the different treatments. Economic and social conditions in the two sets of sites will vary at the outset of the demonstration, and these conditions may change in the course of the demonstration for reasons unrelated to the demonstration itself. In short, both observed and unobserved differences will occur in the initial conditions as well as in the changes in those conditions.

If, at the close of the quasi experiment, ADF outcomes are significantly different from BCE outcomes, the evaluators and the readers of the evaluation must decide whether the difference in outcome is due to the difference in treatment or to one or more of the sources of unobserved difference noted earlier. They may find this decision very difficult (see Hollister and Hill 1995). If there were 1,500 randomized sites, rather than 6, the problem of unobserved random differences across sites would have no practical importance, but experimentation on such a vast scale is unlikely to be attempted.

REASONS FOR CONDUCTING SOCIAL EXPERIMENTS

The motives for undertaking a social experiment invariably differ from one experiment to the next (for a detailed analysis, see Greenberg, Linksz, and Mandell 2003). We begin with the most cynical motive—that the social experiment is conducted to replace difficult political decisions with symbolic action. This charge, which was leveled against the income maintenance experiments, will always have surface appeal for many, legislative and administrative actions have greater symbolism than content. Indeed, some groups often gain a political advantage by delaying a decision, and evaluations with any pretension to scientific standards do take time.

Despite the superficial attractiveness of this argument, we believe such a motive can be easily dismissed. A demonstration may serve as a symbol of a policymaker's sympathies, but no political logic requires a rigorous objective assessment of that demonstration. The usual purpose of delaying tactics is to make the issue go away. If that were the goal, an objective evaluation based on random assignment would act like a bomb with a very slow fuse, detonating years later, perhaps with embarrassing consequences for the policymaker's career.

Policymakers seem well aware of this. Egregiously pork barrel or purely symbolic "demonstrations," or transparently political "programs"—often seem deliberately designed to make future evaluations of any sort impossible. To the cynics, we suggest that policymakers have many other useful pretexts for delay that present fewer risks.

We turn now to less cynical explanations. First, an experiment may be intended for a specific policy decision. Feldman (1989, 80) has contended that if research were intended to influence a specific policy decision, both the timing of the decision and the alternatives to be considered would have to be known in advance. Otherwise, the research would probably not be available when the decision is made or would not be pertinent to the decision. Indeed, these conditions are seldom met in practice.

A less demanding explanation of why social experiments are initiated is that policymakers plan to use the information gained whenever it becomes available. Therefore, if a social experiment demonstrates convincingly that an idea really works, it will help to generate the political support required to place it on the policy agenda.

A final possibility is that the experiment is intended to create an inventory of information for future policymaking (Feldman 1989, 92–96). The
implied intention is for the experiment to contribute to a stock of knowledge, reducing uncertainty should a relevant issue reach the policy agenda.

All but the cynical explanation for social experiments suggest that such demonstrations are intended to generate information relevant to the policy process. Almost without exception, social experiments test the power of particular policies to solve or mitigate serious social problems—long-term welfare receipt; rising health care costs; inadequate or unaffordable shelter for low-income families; long-term unemployment; the clouded future for former offenders, substance addicts, and at-risk youth. Experiments are funded when disagreement over appropriate policy interventions is caused, at least in part, by uncertainty over their potential consequences. If policymakers knew the outcomes, there would be little justification for sponsoring experiments. It is useful to keep this apparently banal idea in mind when considering whether social experiments are ethical and whether they are the best form of research for assessing policy choices.

ARE EXPERIMENTS ETHICAL?

Policy interventions are intended to change people’s lives; it strikes many people as wrong, or at least strange, to change them randomly.

A social experiment is ethical if the treatment received by each group of human subjects is ethical. The experiment is unethical if one or more of the treatments are unethical. The loudest critics of certain experiments have alleged either that the innovation being tested was unethical or that the status quo—the existing laws and regulations governing some aspect of society—was unethical. Random assignment itself is ethically neutral.

In practice, there is usually some restriction on the availability of benefits or services tested (because of budgetary limitation, for instance). If such restrictions exist, then the fact that assignment between treatment and control groups is random is a matter of ethical indifference. A first-come, first-served approach may be more convenient for the program staff but has precisely the same moral value. Shutting the door on people who apply after 4:00 PM Tuesday is just as arbitrary as shutting the door on people who were born on an odd-numbered day, or for whom the last two digits of their Social Security number add up to 12, or for whom a computer generates a random number ending in an even digit.

If we knew in advance that all members of a population would benefit or suffer from the application of a policy toward them, would an experiment on some fraction of them be unethical? Perhaps. Under the principle of horizontal equity, similarly situated individuals should be similarly treated, and singling out some members of a class for arbitrary rewards or punishments not applied to others is, in general, unjust.

When the effects of a policy are unknown, however, the same generalization cannot apply. Many treatments summarized in this book were expected (or at least were intended) to make the individuals subject to them better off but failed to do so; sometimes they made at least some of them worse off. Individuals who were assigned to control groups and, hence, were randomly denied access to those treatments often seem to have lost nothing important. The state of honest ignorance required to justify disparate treatment existed at the time the experiment began.

In other instances, we may believe a priori that members of a target population will be better off, on average, as a result of an innovation (e.g., the income maintenance experiments). However, we may also believe that people outside the target population will be worse off (e.g., taxpayers). We do not know a priori whether the benefits are larger than the costs. Not knowing the trade-off, some policymakers will refuse to institute the innovation until the benefits and costs are better measured, and this is the ethical justification for the experiment.
ALTERNATIVES TO SOCIAL EXPERIMENTATION

Given honest ignorance, it does not follow that random assignment is always the optimal form of evaluation. Social experiments can be quite costly, because they require extensive follow-up data on both treatment and control groups, and some types of policy interventions are better evaluated using none experimental methods. We now examine those methods and consider their advantages and disadvantages relative to experimentation.

The most obvious feasible alternative to random assignment is to compare persons who participate in a particular program with persons who do not participate; the participants then constitute the treatment group and the nonparticipants the comparison group. The least sophisticated version of this approach consists of selecting the comparison group from among people who initially applied for program benefits, but for one reason or another did not actually participate. A more sophisticated version involves drawing a comparison group from among people sampled in national micro-data sets (for example, the Current Population Survey [CPS]) by statistically matching these people on the basis of their personal characteristics with members of the treatment group to make the two groups as comparable as possible.

Other approaches are feasible when a program is carried out in some geographic locations, but not in others. A treatment group, for example, might comprise individuals who live at the sites that have the program, and a comparison group might consist of similar individuals living at other sites. Alternatively, one might compare outcomes in the same set of sites before and after an innovation has been introduced.

An obvious problem with the geographic and chronological comparison groups is that economic and social circumstances differ from place to place and change over time; it is very difficult to control for these factors in a manner that allows the effects of the innovation to be isolated. The superiority of a randomly assigned control group over the "non-participant" comparison group alternatives may be less obvious.

There is essentially one reason for social experimentation: random assignment is the only known means of eliminating selection bias. If an individual can choose whether a policy intervention will or will not apply to himself, then people who choose treatment A will differ from those who choose treatment B in both observable and unobservable ways. The same is true if program administrators make the choice for individuals, rather than having the individuals choose for themselves. Selection factors will also bias comparisons across sites that have and have not implemented an innovation, if, as is usually the case, the decision as to whether to adopt the innovation was made locally. For example, the sites may differ in terms of local economic conditions and in population characteristics.

If individuals in treatment and comparison groups differ only in observable ways, then the analyst can, in principle, adjust for them when researching outcomes. However, absent random assignment, individuals in treatment groups are also likely to differ from those in comparison groups in unobservable ways. The unobserved factors that influence which group they enter are also likely to influence outcomes.

For example, it has been well known since the 1970s that workers who volunteer for training have often experienced a sharp drop in their recent earnings. They cannot properly be compared with workers who have the same observable characteristics but no such dip. Earnings fluctuation may also be permanent or transitory; workers with the same observable dips as training participants will not be comparable if their dips have dissimilar causes.

In one common scenario, workers who have suffered short-term setbacks realize that their careers will recover without additional training, whereas others recognize their old careers have hit a dead end. The latter then try to enter a new path through training, sometimes unsuccessfully. In
comparing workers who have temporary dips and do not enter training with those who have had lasting setbacks and do enter training, we may find that the trained have lower earnings than the untrained. We may then falsely attribute the lower earnings to poor training programs.

In this example, selection bias is negative—it makes the program being evaluated look worse than it is. Selection bias can be either positive or negative, and usually neither policymakers nor analysts know with certainty which is the case. Findings from numerous studies have suggested that selection bias, in whatever direction, can be very large relative to the actual effect of an intervention (Bell et al. 1995; Fraker and Maynard 1987; Friedlander and Robins 1995; Lalonde 1986; Lalonde and Maynard 1987). These studies used the presumably unbiased estimates of program impacts from experiments to assess impact estimates obtained by matching treatment groups with carefully drawn nonexperimental comparison groups. The selection bias associated with the nonexperimental estimates was often larger than the true program impact. Selection bias proved large enough to affect the statistical significance in most cases, the direction of the impact (positive or negative) in far too many cases, and any benefit–cost analysis in nearly all cases.

Bias of unknown sign and unknown but possibly critical magnitude must reduce the usefulness of nonexperimental forms of evaluation. Attempts have been made to develop techniques using nonexperimental data that would have the same degree of internal validity as experimental data (see, for example, Dehejia and Wahba 1995, 2002; Heckman and Hotz 1989; Heckman, Ichimura, and Todd 1997; Heckman et al. 1995, 1998). So far, unfortunately, there is little evidence that these techniques can consistently do so.

For example, Bell et al. (1995) have argued that selection by program staff might generate a comparison group with correctable bias if the criteria for selection are documented and consistently exercised. In principle, the analyst could then control for the observed selection factors. However, the one test of this hypothesis reported in their monograph fails. This might have been because the selection process was not, in fact, fully documented and consistent; this could occur, for example, if program staff selects people who are more likable, attractive, persistent, or literate than those screened out or if the staff selection process effectively allows potential participants to screen themselves in or out.

Most experts in social program evaluation would probably agree with Burtless (1995) and Hollister and Hill (1995) that no alternative method can produce impact estimates that have as much internal validity as a social experiment with random assignment can. This is, however, a reluctant majority. Social experiments are subject to important limitations of their own. Most of these shortcomings concern threats to the external validity of social experiments. The next section lists the most important of these limitations. These shortcomings are not unique to randomized evaluations; they also apply to most forms of nonexperimental evaluation.

THREATS TO EXTERNAL VALIDITY OF IMPACT ESTIMATES

Whether evaluated by random assignment or by nonexperimental methods, innovations to social programs are often tested on a small-scale demonstration or pilot basis. Mankin and Garfinkel (1992) and Garfinkel, Mankin, and Michalopoulos (1992) have suggested that an important component of some policy innovations intended for widespread adaptation is that they cause changes in community attitudes and norms; these, in turn, result in feedback effects that influence the innovation’s success. These authors further suggested that program success depends on information
diffusion to potential participants. They argued that feedback effects and information diffusion will not occur unless the innovation is adopted on a large scale, and that these effects will be missed by small-scale tests.¹¹

Potentially important marketwide effects may also not occur in small-scale tests. For example, training programs could affect the level of wages in a community. They might also affect the number of employers in a community, if enough workers receive the training to induce firms to move into the area. Little is usually known about the importance of such effects. One can only speculate whether small-scale tests of social policies are seriously biased by their absence.

In arguments that are also applicable to non-experimental evaluations of small-scale demonstrations, Heckman (1992), Heckman and Smith (1995), and Manki (1995) have contended that participants in small-scale experiments may not be representative of individuals who would participate in ongoing, full-scale programs. This could occur because of a lack of information diffusion, the reluctance of some individuals to subject themselves to random assignment, resource constraints in full-scale programs that result in program administrators restricting participants to people meeting certain criteria, and numerous other reasons.¹²

One approach for eliminating biases caused by testing innovations on a small scale is to incorporate them on a sitewide basis in some locations and use other sites (perhaps statistically matched) that have not adopted the innovation for comparison purposes. For example, the effect of housing subsidy on rents in a housing market can only be learned through “saturation” of the community—that is, by providing the subsidy to all eligible households (see Lowry 1983). A saturation design, however, does not allow feedback, information diffusion, and market effects to be measured separately from other types of program impacts. Moreover, as previously discussed, such a design will produce biased impact estimates if the treatment and comparison sites differ in ways that are inadequately controlled for in the evaluation.

Social experiments are always limited to relatively few geographic areas. This is also true of many non-experimental evaluations, especially those that rely on saturation designs. Because these sites are rarely selected randomly,¹³ the external validity of the evaluations can be questioned (see Heckman 1992; Heckman and Smith 1995; Hotz 1992). Difficulties in obtaining a representative sample of program sites are especially acute in cases where local administrators’ cooperation is essential. However, the degree to which site selectivity translates into bias in the results of an impact analysis has not been proved empirically.¹⁴

Another potential shortcoming of social experiments concerns entry effects (Manski and Garfinkel 1992; Moffitt 1996). For example, if only unemployed people or people with incomes below certain thresholds are eligible for a training program being evaluated experimentally, and the services provided by this program are perceived as beneficial, some ineligibles may leave their jobs or otherwise reduce their incomes to qualify. In welfare-to-work programs that are mandatory for transfer recipients, and perceived by them as burdensome, some individuals who might otherwise have entered the welfare rolls may decide not to do so to avoid participating. Measuring entry effects in a random-assignment context requires that ineligibles be included in the evaluation sample, but because of cost considerations, this is seldom done.¹⁵

A few experiments stand out as exceptions to this rule. In the Self-Sufficiency Project evaluation, a random assignment experiment recently conducted in Canada (Michalopoulos et al. 2002), only people who had been on welfare for at least one year qualified for the substantial increase in transfer payments made available through the tested program. By design, the experimental sample included transfer recipients who had been on welfare for less than one year—to see if some of them would extend their stay on the rolls to qualify for the new benefit schedule (Berlin et al. 1998). Similarly, the income maintenance experiments included households whose earnings at
intake were too high to qualify for program benefits—to see if some of these households would reduce their earnings to qualify.

The behavior of participants in a pilot test of program or policy could be influenced by knowledge that they are part of the pilot test, a so-called “Hawthorne effect.” For example, if members of a research sample knew that their behavior will be measured in terms of certain outcomes, such as earnings or educational attainment, some of them might attempt to succeed in terms of these outcomes. There is virtually no information about whether Hawthorne effects bias findings from social experiments. It seems possible that members of both the treatment and control groups could respond similarly to being part of a social experiment. If so, such effects will cancel out in measuring impacts, and there would be no bias. Alternatively, some control group members could be discouraged by the fact that they were allocated to the control group, rather than the treatment group, and alter their behavior for that reason.

In both experimental and nonexperimental evaluations, members of the control or comparison group may, in practice, receive many of the same services as those received by members of the treatment group. For example, in the case of training programs, training of many types may already be available through community colleges and adult schools, and members of control or comparison groups can often obtain financing for these activities through existing nondemonstration sources. Consequently, estimates of program effects on participants do not measure impacts of the receipt of service against the nonreceipt of services. Rather, such estimates represent the incremental effect of the program services over the control services.

The existence of alternative services in the community and the measurement of incremental effects are not necessarily detrimental to an evaluation, depending on the evaluation’s goal. Often the goal is to determine the effect of an innovation relative to the actual environment, rather than relative to the complete absence of alternative services. Nonetheless, in interpreting impact estimates from social experiments and other evaluations, one should keep in mind that these estimates usually pertain to incremental effects, rather than to a control environment in which no services are available. Naturally, this limits the extent to which findings may be generalized from one environment to another. In addition, what Heckman and Smith (1995) call “substitution bias” is possible. Substitution bias can occur if a new program being tested experimentally absorbs resources that would otherwise be available to members of the control group or, instead, if as result of serving some members of a target group, the new program frees up resources available under other programs that can now be used to better serve member of the control group.

The summaries in this Digest give the reader some sense of the ambiguities that even the best experiments can create. We have also included some outright failures, where the demonstration never answered and, in hindsight, never could have answered the policy questions asked. Social experimentation is a specialized research tool, suitable for some inquiries and not for others—a chisel that cannot substitute for a screwdriver, but is the best instrument yet developed for many purposes.

**ARE EXPERIMENTS UNDULY EXPENSIVE?**

Social experiments have also been criticized for their cost (Levitan 1992). Experimental costs fall into three basic categories:

1. Implementing the innovative treatments;
2. Collecting data on the different groups; and
3. Analyzing the data.

It is worth noting that most of these costs are typically incurred regardless of whether a program is evaluated through random assignment or using nonexperimental methods. The major difference in costs is that the random assignment process
must be carefully implemented and monitored, but this is usually not very costly. In addition, the data needed for a nonexperimental evaluation (but not a random assignment evaluation) is occasionally already being collected for other purposes. If so, a nonexperimental evaluation can be considerably less expensive than one that relies on random assignment. However, it is fairly rare when the data needed for an evaluation do not have to be specifically collected for that purpose; and when this does occur, it is usually in the case of an evaluation of an ongoing program, rather than a test of a new policy initiative. Most social experiments have been used to evaluate new policy initiatives.

There are two perspectives on the costs of a social experiment. The first compares costs with benefits: Will the experiment be likely to generate information sufficiently valuable to the political or institutional process to justify the costs? There are three facets to this question: it must be answered prospectively, before the information exists; it involves the value of information provided by social scientists in a democracy; and it must be explicitly answered affirmatively by decisionmakers before the experiment can take place.

A somewhat different perspective looks at the experiment’s opportunity cost, weighing the net benefits of the experiment against the net benefits of the most probable alternative use of the resources. If the funds are federal, the most probable alternative use is another project at the same agency or agencies.18 Perhaps, for example, the funds would be better spent on

- Implementing a nationally representative in-depth survey, of, say, 50,000 households (at an annual cost in the United States of around $20 million, with the exact cost depending on design choices);
- Demonstrating the feasibility of implementing a program change, rather than (or at least prior to) attempting to measure its impact (at an evaluation cost of between $20,000 and $20 million, depending on scale and number of sites); or
- Improving program management information systems (costs vary too much to provide a range).

From either perspective, sponsoring a social experiment requires complex resource allocation decisions. A variety of different decisionmakers (including politicians, political appointees, civil servants, and foundation directors), representing a wide spectrum of political views, authorized the social experiments conducted to date. Some were professional social scientists who could readily evaluate the technical merit of a proposed experiment on the basis of their own training and experience; the rest usually had as much access to expert opinion as they wished. It is striking that many very different individuals decided that this type of investigation is worth its costs. Nevertheless, controversy over the use of experimental techniques continues.

WHAT “OPTIONAL FEATURES” CAN EXPERIMENTS HAVE?

All social experiments are intended to provide impact estimates of the tested policy or policies. These estimates have already been discussed at some length. In addition to impact analyses, social experiments also commonly, but far from universally, feature two other types of analyses.

One of these, “process analysis,” goes by sundry other names and has a number of purposes. For example, an experiment’s sponsor may desire third-party verification that the treatments were administered as planned. The sponsor may wish to know how many people received the treatment and the nature and intensity of the services received. The sponsor may want to know the character of the environment in which the services were delivered and whether subjects understood the incentives provided to them by the experiment. The process analysis may also attempt to convey
participants' and the program staff's subjective reactions to the experiment, or may speculate about whether certain identifiable subgroups would be likely to experience a greater or lesser impact than the sample as a whole.

An experiment without a process analysis is not necessarily useless. One may test the performance of a machine without any knowledge of the internal components. The lack of attention to "how it all really works" need not invalidate the findings, but analysts may have greater difficulty interpreting the data and may be more prone to err in attempting to do so. Lack of attention to nonexperimental factors can lead to a "type-two" statistical error—the finding of no policy impact when in fact an impact exists. For one thing, a process analysis may reveal that an innovation was never really implemented, and therefore data analysis is (probably) pointless; or that the assignments were not really random, so that internal validity is questionable.

A more subtle contribution of process analysis is in discovering relevant subgroups. These may only become clear to people who spend time in the field observing and interviewing. If two groups in an experiment really do differ only randomly, then a simple test comparing the mean outcome in one group with the mean outcome in the other yields an unbiased estimate of policy impact. Analysis by subgroup, however, may show that the treatment has significant impact for subjects of one type but not another. For the same reason, process analysis may turn up nonexperimental variables that should be controlled for in a regression analysis.

A second common, but far from universal, feature of social experiments is a benefit-cost analysis. Even policies with positive benefits for the target population are unjustified, if the cost to the rest of society is too high. A benefit-cost analysis aggregates benefits of the tested policy over time, both to participants and to society at large, and compares them to costs. Such analyses have greater value for some demonstrations than others. There is no need to compare benefits with costs, for example, if there are no demonstrable benefits.

Benefit-cost analyses encounter numerous difficult problems; what value to place on the loss of leisure to participants in training and welfare-to-work programs; how much future benefits and costs should be discounted; how to extrapolate experimental impacts beyond the period over which data were collected. These are all areas of significant controversy. The value to nonparticipants of changes in the behavior of program participants may be especially difficult to quantify. There is little evidence, for example, on the dollar value that nonparticipants place on seeing welfare recipients go to work; or on low-income people enjoying increased consumption of health care, food, or housing; or on reductions in criminal offenses.

HOW MIGHT SOCIAL EXPERIMENTS BE USED?

Frequently, it is anticipated that results from a given social experiment will lead to a yes or no decision on the policy being tested. However, the relationship between experimental findings and policy decisions often does not appear to be so direct. Findings from social experiments are used in many ways, some of them unanticipated.

For example, it is unlikely that findings from the income maintenance experiments had any major role in the failure to adopt the policy tested, the negative income tax (Greenberg, Links, and Mandell 2003, chapter 6). However, the experiments did demonstrate that certain innovations used in administering the transfer programs (monthly reporting and retrospective accounting) could be successfully implemented. Partly as a result, these innovations were adopted nationally in existing welfare programs.

Findings from the income maintenance experiments also altered the commonly held pre-experimental view that extending cash assistance
to intact families would enhance their marital stability. The experiments suggested that this did not occur. In addition, the income maintenance experiments provided useful information about the effects of transfer programs on hours of work by different types of adults. These findings were in turn used to guide decisions about other transfer programs, such as Food Stamps and Aid to Families with Dependent Children.

There are at least three different dimensions on which using findings from an experiment (or any other types of evaluation) might be mapped (Greenberg, Linksz, and Mandell 2003). First, findings may either influence specific policy decisions or address unresolved scientific or intellectual issues. In our previous example, the finding that monthly reporting and retrospective accounting were operationally feasible affected the everyday administration of welfare programs, a policy effect. Both the unexpected findings on marital stability and the expected findings on the relative inelasticity of male labor supply had broader intellectual impacts.

Second, findings used may be more or less central to social policy. On one end of the continuum, research findings may influence core policy decisions or general intellectual orientations; at the other end are cases in which research findings influence relatively narrow elements of policy and its implementation—elaborative or peripheral uses. The income maintenance experiments demonstrated that male heads of household did not cut back much on hours of work when an income guarantee was available; this finding was a major intellectual contribution. The same experiments showed that retrospective accounting was feasible; this finding was elaborative or peripheral (which is not the same as unimportant).

A final dimension of utilization distinguishes between prededication and postdecision utilization (see Majone 1989, chapter 2). An old joke informs us that some people use statistics the way a drunk uses a lamppost—not for illumination but for support. If the positions of policymakers are established, at least in part, on the basis of research findings, we can say the research was used for illumination (the use is prededication). Postdecision utilization refers to the use of research findings to advocate already established positions. Findings on retrospective accounting seem to have changed the minds of key administrators; findings on male labor supply seem to have changed nobody’s mind, but were used to support previously determined positions (see Greenberg, Linksz, and Mandell 2003, chapter 6, for further discussion).

**WHAT IS THE HISTORY OF SOCIAL EXPERIMENTS?**

The idea of a control group was firmly established by 19th-century pioneers of medical and biological research. For example, in one classic experiment, Louis Pasteur divided a flock of sheep into two groups. In one group he injected attenuated material from other animals that had died of anthrax, so that they could develop immunity; the second group did not receive this injection. Both groups were then injected with anthrax-infected matter. None of the treatment sheep became sick, but all the control sheep died.

If Pasteur’s results had been less dramatic, critics would surely have claimed that for some reason the control group was more anthrax-vulnerable than the treatment group. Over time, the charge that controls were “inadequate” became commonplace, without any rigorous idea of adequacy being developed, since no treatment group was ever identical to the control group.

The concept of randomization, like many other fundamental statistical tools, was conceived by Ronald Fisher.21 The concept appeared initially in a 1925 book by Fisher, Statistical Methods for Research Workers, and then was fully elaborated in his 1935 book, The Design of Experiments. Fisher pointed out that no two groups could ever be identical because every organism, test tube, soil
sample, and so forth would vary slightly from every other. Therefore, the researcher's task was to design the difference between groups in such a manner that the mechanism for allocating cases to one group or another could not be related to the issue being studied. Allocation by pure chance (a coin flip, a table of random numbers) did exactly that.

Social experimentation, however, did not come about until much later. It is usually traced to the New Jersey Income Maintenance Experiment, which was initiated in 1968. The idea of conducting an income maintenance experiment is attributed to Heather Ross, who in 1966 was a Massachusetts Institute of Technology graduate student in economics. In that year, Ross was beginning work on her dissertation as a fellow at the Brookings Institution in Washington, D.C. Ross was frustrated that inferences about the responses of low-income people to transfer payments could not be readily drawn from existing data. She was also concerned by the use of unsubstantiated anecdotes about welfare recipients by politicians. She wished to collect data that could be used to determine what poor people would actually do if they were provided money. Would they work less? Would they quit work altogether? How would they spend the additional money? To answer such questions, she proposed to conduct a random assignment experiment.

To fund this project, Ross wrote a proposal in 1967 to the U.S. Office of Economic Opportunity, which had a staff of social scientists. Ultimately, as Ross puts it, she ended up with a "$5 million thesis," which at the time was an extraordinary sum of money for a single social science research project.

Her proposal germinated in the New Jersey Experiment. Its importance as a landmark in social science research is hard to overstate. Although the technique of randomly assigning individuals for purposes of clinical health trials and educational innovations had been used for years, the New Jersey Experiment was the first prominent study to use this technique to test social programs. Other social experiments, including additional income maintenance experiments, followed fairly quickly.

**WHAT IS THE FUTURE FOR SOCIAL EXPERIMENTS?**

Numerous experiments are still in progress, and therefore are only briefly summarized in this Digest. We expect many more to be initiated in the future. Intractable social problems, unfortunately, show no sign of vanishing; substantial uncertainty clouds debate over proposed solutions; and the selection problem remains.

However, certain types of social experiments will probably not be conducted in our lifetime, even when technically feasible. One such area is community development, where the appropriate unit of analysis is neither the individual nor the household, but the neighborhood or city.

An example illustrates. Suppose it is hypothesized that a $5 million investment of a particular type will generate more than $5 million in benefits in an average community of 10,000 people. However, the benefits are expected to be sufficiently modest—say $5.3 million—that 250 treatment communities and 250 control communities would be required in order to have an 80 percent probability of detecting the benefits with 95 percent confidence. Therefore, testing the hypothesis requires distributing a total of $1.25 billion to 250 randomly selected communities throughout the country. Such a test would allow one community to enjoy a benefit denied another, not because of some formula in law, and not because of the personal and political influence of their representatives in Congress, but because of the impersonal operation of a lottery. The aggregate appropriation and the number of communities not funded would make the program difficult to ignore or to gloss over. It seems highly unlikely that any Congress would countenance this experiment or anything like it.

Other policy controversies do not seem technically amenable to the experimental evaluation. A topical example might be the sensitivity of taxpayer
savings and investment behavior to the tax rate on capital gains. Suppose one group of taxpayers were randomly assigned to a capital gains tax rate lower than the rate for other taxpayers, in order to evaluate the effects of rate reduction on the national economy. Members of the treatment group could obtain arbitrage profits by getting other taxpayers to sell property to them at below-market prices, permitting the sellers to avoid (evade) much of the capital gains tax that would otherwise be levied on them. (The arbitrage profit would be split with the sellers through a side payment.) Obviously, this would bias the estimated impact of the experiment.

The general point is that the feasibility of the experiment may depend on the ingenuity and deviousness of the target population. On the other hand, the people who design and implement the experiment may also have ingenuity, expertise, resources, and perhaps a little deviousness of their own. In the previous example, the detection of tax arbitrage activity is a constant pursuit of the tax authorities; we merely doubt that they would be willing to devote the substantial resources that an adequate capital gains tax experiment would require.

**USES AND ORGANIZATION OF THIS BOOK**

This volume is intended to serve as an archive, a reference, an "armory," and a textbook supplement:

- **Archive.** Social experiments deliver credible evidence on the responsiveness of human behavior to policy, but the findings from most social experiments have been only narrowly disseminated. We want to increase awareness of and access to this material, much of which may eventually become inaccessible in its current form.

- **Reference.** We hope to provide a one-stop guide to all experimental findings on particular interventions, which can be used as a first step in developing literature reviews, options papers, and the like.

- **Armory.** The idea and practice of social policy experiments remain controversial. We expect that this book will supply all sides of the debate with weapons and ammunition.

- **Textbook supplement.** We hope that these experiments can give students in a variety of courses—public policy and administration, economics, social work, education, vocational rehabilitation, sociology, statistics, political science, metropolitan planning, and other fields—an appreciation for the interaction of theory, policy, statistics, and daily life.

To develop the list of social experiments summarized in this volume, we began with the 37 compiled by Greenberg and Robbins (1986). We learned of a few additional older social experiments from a lengthy bibliography by Boruch, McSweeney, and Soderstrom (1978). About a dozen experiments in the early childhood intervention literature came from a literature review by Benasich, Brooks-Gunn, and Clewell (1992). Names of additional experiments were obtained in response to numerous written, telephone, and e-mail requests to academics, government employees, and employees of prominent social science research firms. Others were listed in abstracts of research reports provided on the web by the Economic Research Network (ERN). Still other experiments were acquired from various reports disseminated by research firms and the government and from journal articles. Finally, we used "snowballing" techniques: when interviewing people associated with experiments with which we were familiar, we asked if they knew of any others. We may have overlooked some smaller social experiments, but we do not believe we have missed many large ones.

The first two editions of the Digest listed only four social experiments conducted outside the United States. For this edition, we have made a
special effort to locate such experiments. We have added 29 completed experiments, six completed quasi experiments, and three ongoing experiments to the four in the previous edition. We missed 18 of these in assembling the previous edition, while the remainder were initiated after work on that edition was completed. Although we are still more likely to miss experiments conducted outside, than those conducted within, the United States, it seems apparent that the United States accounts for the vast majority of social experiments. Twenty-five of the experiments listed as "ongoing" in the second edition of the Digest were not completed and, consequently, are not included in the third edition. In 23 of these planned experiments, either the planned treatment was not implemented or (much more often) it was implemented but not formally evaluated. In the remaining two cases, formal evaluations were conducted, but the random assignment design was dropped. All but two of the 25 planned experiments were targeted at public assistance recipients. Almost all of those targeted at public assistance recipients were terminated in response to passage of the Personal Responsibility and Work Opportunity Reconciliation Act in 1996. In most instances, states had been required to conduct experimental tests of welfare reform provisions that they wanted to implement in exchange for receiving federal waivers; but the need for these waivers was removed by the 1996 legislation.

The information contained in the experiment summaries came from two major sources. First, we reviewed at least one research report on each experiment. Second, information not available from the research reports was obtained from telephone interviews and e-mail exchanges with staff members of the organizations that conducted the evaluations and the government agencies that sponsored them. Often one interaction of this sort was sufficient, but sometimes several were necessary.

Most items appearing in the summaries for each social experiment are self-explanatory, but a few require a brief comment.

Cost: Of all the information about social experiments that we attempted to collect, cost data were the most difficult, for a number of reasons. Sometimes the information once existed, but the necessary records could no longer be located. When an experiment was administered by an existing government agency, it was often difficult to separate the incremental cost of administering the experimental treatment from other costs incurred by the agency. Sometimes the total cost was available, but administrative costs could not be separated from the evaluation costs. In other cases, the experiment was part of a larger research project, and costs were not separately allocated to the experimental and nonexperimental parts of the research. In the summaries, we usually attempt to indicate what the cost figures we report cover. Costs shown are contemporaneous with the demonstration; we do not adjust them for inflation and seldom convert foreign currency to dollars.

Number of Treatment Groups: The number of treatment groups always includes the group or groups used for control purposes.

Major Findings: Information on findings reported in the summaries was typically obtained from the final reports we reviewed. For a few experiments, alternative sets of findings have been produced by a methodological approach different from that used in the final report. We usually ignored such findings. Some experiments have a large volume of results. To keep the summaries brief and the reported numbers manageable, we have concentrated on those findings that pertain as directly as possible to the major experimental outcomes of interest.

Design Issues: Numerous issues arise in selecting treatment and control groups, administering experimental treatments, collecting data on outcomes, analyzing these data, and so forth. Decisions concerning these issues can have a major influence on findings from social experiments, sometimes rendering the findings invalid. For each experiment, we attempt to alert readers to the most important
design issues. The number of design issues we list for a particular experiment is unimportant. Some design issues are much more important than others. Although we have uncovered some of the design issues we mention on our own, in many instances we simply repeat information we found in the evaluation reports. Careful evaluators are especially likely to list important caveats.

**TREATMENT ADMINISTRATION:** The treatment administrator is the organization responsible for rendering the treatment. Treatments tested in early social experiments were often administered by separate offices set up by the evaluator expressly for that purpose; since 1975, the treatments tested in most social experiments have been administered through existing agencies.

**ENABLING LEGISLATION:** Federal and state legislation have often mandated that social programs, especially those tried on a pilot basis, are evaluated. Some experiments that we reviewed began as a result of legislative mandates for evaluation.

**POLICY EFFECTS:** In a number of instances an experiment’s findings appear to have directly influenced the direction of policy, and we say so. As with all other elements of the summary, we admit the possibility of errors of omission and commission in these conclusions.

**INFORMATION SOURCES:** Each summary indicates at least one source for more detailed written information (usually a final report) on the experiment described. In some cases, the report has been published in a journal or book, but more typically it can only be obtained from the research firm that conducted the evaluation or the government agency that sponsored it. In a few instances, the report is not written in English.

We have sorted the experiments by their target populations: welfare recipients, unemployed, and so on. Within each target population, experiments appear in approximately chronological order. Thus, brief summaries of ongoing experiments appear after the longer summaries of completed experiments.

Following the summaries of the experiments, an article entitled “The Social Experiment Market” appears. This article, which was originally published in the *Journal of Economic Perspectives* in 1999, is largely based on information extracted from the summaries of social experiments included in the second edition of the *Digest*. The article describes the characteristics of social experiments and trends in social experimentation. We reprint the article here because it appears that few of these characteristics and trends have changed in major ways since the second edition was published in 1997, even though a large number of social experiments have been initiated since then. The important changes that have occurred since 1997 are discussed in a brief postscript to the article.

The volume contains two indexes. The first lists individuals and organizations that were either associated with particular experiments or acted or commented on these experiments. The second index references the policy interventions tested by the experiments summarized in this book.

**Notes**

1. Two recent books that do this are Boruch (1997) and Orr (1999). In addition, in a series of essays, Hausman and Wise (1985) also discuss specific social experiments and issues that have arisen in conducting them.
2. If the behavior of a larger randomized unit (e.g., a local government, a business) is itself the subject of analysis we consider the demonstration an experiment. If the aggregated behavior of individuals associated with that larger unit is the subject of analysis we consider the demonstration a quasi experiment. Our rule is that experiments analyze and randomize at the same level of analysis, while quasi experiments analyze at a lower level than they randomize.
3. Moreover, people may move from site A to site B and vice versa, during the demonstration, and these moves may or may not be related to differences in the treatment. If people move because of differences in the treatment, then the validity of the evaluation will probably be compromised.
4. Observed differences across sites typically cause less serious problems than unobserved differences because they often can be controlled for statistically.
5. In the specific case of the income maintenance experiments, the fact of an ongoing income maintenance
experiment was not used to delay action. On the contrary, premature, and eventually contradicted, data from the experiment were used in an attempt to push forward an income maintenance proposal by the Nixon administration that was similar to the one tested.

6. In poverty programs, some would argue that the social welfare evaluation is more complex; perhaps current benefits do long-range harm to recipients, whereas the reduction of poverty may raise the well-being of nonpoor taxpayers. These concerns are valid, but are tangential to this issue.

7. As previously discussed, quasi-experiments involve using random processes to assign sites to treatment or control status. However, quasi-experiments of this sort are relatively rare. One reason is that program administrators at local sites are generally very reluctant to relinquish their prerogative to decide whether to adopt particular innovations.

8. In practice, even observable differences may interact nonlinearly with treatment, and the analyst may fail to adjust for them correctly.

9. Burtless (1995) assessed most of the problems associated with conducting social experiments and vigorously defended the usefulness of the technique.

10. There is no inherent reason why social experiments must always be small-scale. For example, in some experiments, all members of the target population within a particular geographic area, but a small group of randomly selected controls, are eligible to receive program services.

11. Whereas the lack of feedback and information diffusion can threaten external validity, the presence of such effects might threaten internal validity. One possibility is that widespread media coverage of an innovative treatment might mislead control group members about what rules affect them. For example, our summary of the New Jersey Family Development Plan notes a controversy among analysts about whether subjects’ confusion might have affected outcomes in that demonstration. A relevant question is how the concept of “confusion” is to be measured.

12. Note that the main problem is whether small-scale demonstration results generalize to large populations. Randomization per se is a relatively unimportant element here. If the pilot program has limited slots, the target population is necessarily self-selected in applying and not applying, regardless of whether the allocation mechanism is random or nonrandom.

13. Note that this refers to selecting sites randomly from the population of potential sites, not to a saturation design in which sites are assigned randomly to program and control status.

14. The standard argument is that only sites operating superior programs will acquiesce to an evaluation. However, there may be only minimal correlation between local operators’ self-appraisals and the results of a rigorous third-party evaluation. Indeed, even when sites are self-selected, estimated impacts are typically modest, suggesting that any site selection bias is not large enough to lead to an unwarranted expansion of a program in response to inflated impact estimates. Nevertheless, biases of unknown magnitude may arise from management differences among self-selected sites.

15. In nonexperimental evaluations, it is sometimes possible to estimate entry effects by using a site saturation design (Johnson, Klepinger, and Dong 1994; Marcotte 1992; Schiller and Brasher 1993; Wolf 1990).

16. We report some information about costs in the summaries of the individual experiments.

17. Costs associated with administering innovative treatments are typically counted as part of total experimental costs when the experiment is administered through special offices set up expressly for the purpose, but are rarely counted if the experiment is administered through an existing government agency.

18. The cost of the experiment will typically be a fraction of the rounding error in the appropriation for any major program of the agency, but this does not justify suboptimal use of the agency’s research budget.

19. The more one knows about bureaucracy, the less surprising this news would be.

20. The most potentially damaging findings came from the Seattle-Denver Income Maintenance Experiment, which seemed to indicate that a negative income tax program would cause family splitting. When these findings first became available, the Carter administration was attempting to promote a negative-income-tax-like program in Congress. The only member of Congress known to have reacted strongly to the findings was Senator Daniel Patrick Moynihan of New York, who said: “We were wrong about a guaranteed income! Seemingly it is calamitous!” (quoted in Demkovitch 1978). Although Moynihan was an influential authority on welfare, for many other reasons the administration’s legislation never got out of committee in the U.S. House of Representatives and thus never reached the Senate. It is not clear that even Moynihan’s opposition was primarily motivated by marital stability findings; the driving force may have been fiscal relief for New York State. See Greenberg, Linske, and Mandell (2003, chapter 6) for greater detail.

21. Randomization also seems to have been conceivably independent of Fisher, and slightly earlier, by W. A. McCall (1923).

22. The information in this paragraph is based on a telephone interview with Heather Ross and has been con-
firmed by several other people who were involved in social experimentation in its early stages.

23. For classic early statements of the rationale, see Orcutt and Orcutt (1968) and Rifkin (1971, 86–119).


References


