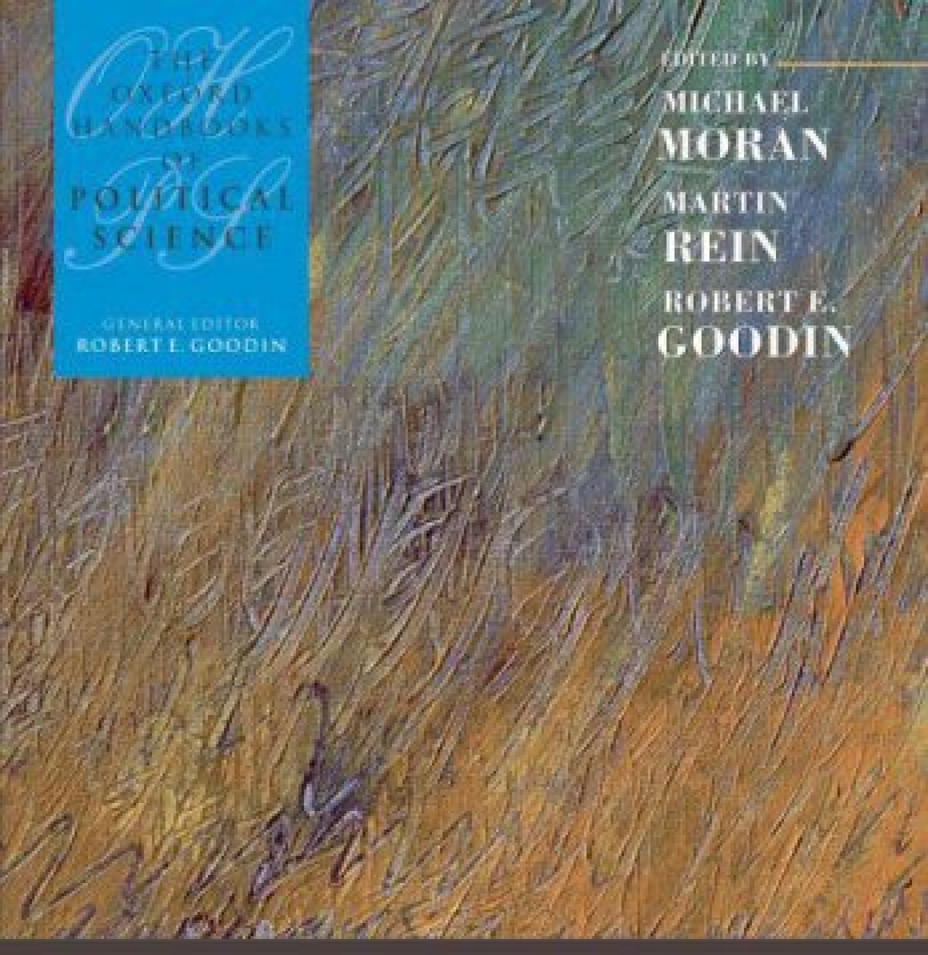

THE OXFORD HANDBOOKS OF POLITICAL SCIENCE

GENERAL EDITOR
ROBERT E. GOODIN

EDITED BY

MICHAEL
MORAN

MARTIN
REIN

ROBERT E.
GOODIN

The Oxford Handbook *of* PUBLIC POLICY

6. Limitations of SEs

6.1 Policy Limitations

Effects on Decisions

When we review the history of social experiments, we see that they have not had a decisive, direct effect on the ensuing decisions. Of our four examples, only the welfare-to-work experiments were later reflected in policy. Neither the health insurance experiment, the nursing home incentive reimbursement experiment, nor the income maintenance experiments made much of a dent at all, and the findings were relegated to the great analytical storehouse. Even in the welfare-to-work experiments, where experiment results seemed to affect later policy, the result was at best indirect.

Greenberg, Mandell, and their colleagues did a telephone interview study of welfare directors in the states. They found that while most of the state directors knew something about the findings of the welfare-to-work experiments (although not the specifics), they didn't believe the findings had influenced the policies of their own state. What they did value was the demonstration that states could administer the program without much problem and a general sense that work first was better than training first for former welfare recipients. In their 2003 book, Greenberg et al. conclude:

> Ironically, however, even though these experiments did have important effects on policy, their role was nonetheless limited . . . In particular, many policymakers already viewed the programs tested by the welfare to work experiments as attractive on other grounds. Findings from the experiments simply reinforced that view. Consequently, rather than being pivotal to whether the types of programs they tested were adopted, they were instead used persuasively and in designing these programs. In other words, they aided policymakers in doing what they already wanted to do. (2003, 308, 310)

Why should the results of SEs be so marginal? Why doesn't rationality reign?

Social scientists are under no illusions that "scientific evidence" will displace all other sources of understanding. Policy making is also based on ideologies and beliefs, interests, competing information, and institutional norms (Weiss 1983, 1995). The results of social experiments can nudge policy only a small distance, and their influence is dependent in large part on the interplay with the other factors in the policy environment. Social scientists know that legislators and administrative officials have long-standing beliefs and principles that guide much of their orientation toward policy. Their ideological orientation exerts powerful influence over which policy proposals receive even a hearing. Attitudes toward abortion and gay marriage are obviously determined by ideology and principles, but it is not only on such extreme issues that ideology often prevails. For some policy makers, similarly strong beliefs affect their views of the enactment of a draft, the need for standardized performance tests in schools, mandatory sentences for repeat offenders, and needle exchange programs for drug addicts.

Interests are always powerful influences on policy. Drug manufacturers, farmers, radio station owners, state and city service workers, trial lawyers, charities, utility companies, universities, hospitals—almost every organized body in the nation seeks to promote its own well-being through public policy. The jostling among organized interests provides much of the drama in the policy arena. The scene is marked by the formation and dissolution of temporary coalitions of interests as the issues on the agenda shift and change.

Nor does social science represent the only form of legitimate information. The policy world is awash with formation. Lobbyists hawk their own version of past events and futures. Media columnists and editorial writers add to the stew. Many organizations have their own in-house information resources—databases, research units, news services. The availability of 24/7 web-based information in titanic proportions makes getting information much less difficult than interpreting the information with a sense of history and context.

Furthermore, each institution in the policy system has its own set of rules and norms. The US Congress, for example, proceeds according to a system of committee appointments, minority/majority representation on committees, vote taking, reporting to the full body, closing off debate, reconciling different versions of bills passed by the two houses, as well as time schedules, budget limits, pressure group access, and so on, that have major influence on the nature of policy that emerges. Ron Haskins (1991) tracked the instances that the MDRC research was mentioned at various times in the welfare reform policy process and found fewer and fewer specific mentions of the MDRC research as the welfare policy made its way through hearings, bill writing, and consideration in the House and finally in the House–Senate Conference. The internal norms and culture of each institution in the policy system exercise great pressure on its own activities and on the activities of other institutions with which it interacts. These four sets of influences—ideology and beliefs, interests, other information, and institutional norms—set limits to what social science can contribute and how much attention it can mobilize. Social experimentation, as one small subset of social science research, is even further constrained by the surround.

Misuse of Research Findings

The results of SEs can be misused in policy discussions (Orr 1998). As with any source of information, policy makers may choose to disregard results if they are not congruent with their own beliefs and political agendas. During the congressional welfare reform debates, the welfare-to-work research was used to argue that education and training were effective strategies and that large amounts of federal funding were needed to produce effects. In fact, education and training received little attention in the programs studied, and the experiments showed that relatively low-cost job search and work experience were effective (Haskins 1991).

Policy makers may take note of the general public reaction. If the public is not interested or is skeptical of certain results, policy makers have little incentive to push forward any change based on the results. Results may not even reach the ears of policy makers if the sponsoring agents of the studies themselves do not like the results. What goes to publication can be influenced by the satisfaction (or dissatisfaction) of the agency that asked and paid for the study in the first place. Less insidious is a simple lack of dissemination of experiments' results. In the nursing home incentive study, the departure of the federal staffer who had sponsored the study contributed to the lack of dissemination of the findings. Few people learned of the results, and little use was made of the findings (Greenberg et al. 2003). A reanalysis of the data that showed more positive results from incentives (Norton 1992) went almost totally unnoticed.

Contributing to the risk of misinterpretation or misuse, policy makers may not have a particularly honed sense for the quality of research or indeed have the skills to interpret results correctly when they are presented with them (they are not alone . . . it is difficult for everyone). Policy makers tend to rely on indirect indicators of quality such as the reputation of the researchers, how the research community reacts to the results, and whether the research fits with their own preconceived notions of what the results should be (Orr 1998).

Simplistic Thinking

SE encourages policy makers to ask a simple question: What works? It leads them to think that social scientists can identify one policy that has the desired results. It discourages them from asking follow-up questions: For whom does it work? Under what conditions? What kind of implementation is necessary? How much difference does it make? What are other alternatives and how effective are they?

Ability of Researchers to Work in the Policy World

Social experiments take place in the messy world. The kinds of social scientists who have the requisite knowledge of research design, sampling, measurement, and statistical analysis are not always the kinds of social scientists who communicate well with political actors. Experimenters in these circumstances have to listen. They have to be aware of what policy options are feasible. They should know the history of political battles already waged on the turf. And still they have to know the scientific literature and the intricacies of research design and conduct. Such people can be hard to find. In their stead come highly skilled researchers who may have little skill, and often less interest in aligning their experiment with the world of politics.

Heightened Scrutiny

The results of social experiments may fare somewhat better than other research findings as they are less assailable by opponents. This occurs, in part, as the research community tends to support the results of randomized experiments and thus, may present a more unified front for policy makers trying to understand what researchers believe. Thus, for example, the health insurance experiment produced generalized agreement among the research community that cost sharing could reduce health care without detrimental effects on health—a question that until then no study had adequately answered. And yet, even some of the best social experiments are open to methodological critique and indeed sometimes may be treated to a more rigorous critique than might be expected due to their high visibility in both the research and the policy worlds. The school choice experiments are an example (e.g. Howell and Peterson 2004; Krueger and Zhu 2004). Because parental choice of schools is such a politically loaded issue, studies are scrutinized in meticulous detail.

6.2 Research Limitations

Social experiments are not easy to bring off. To be at all persuasive, social experiments require big slugs of time, lots of money, powerful research expertise, and enough flexibility to respond to changing conditions and questions while the experiment is in process. The impact of social experiments on policy making is limited not only by the political process but also by the constraints and limitations of the research world. Social science methods themselves are not always ideal for describing and analyzing complex policy issues.

Design Challenges

Researchers are plagued by a series of challenges when conducting research in the real world. Experiments pose difficulties all along the way. The first problem is choice of sites. Even though the policy option that an experiment is testing is usually intended to apply to all members of the relevant group in the nation (or the state), the experiment cannot be implemented among a random sample chosen throughout the nation. The intervention can be offered (and studied) in only a few places. Even the most expensive SEs have had to limit the intervention to a few sites. How does the researcher decide what sites are “typical” or “representative” enough to stand in for the nation as a whole? Researchers avoid places with obviously unusual features, but much of the choice depends on which sites agree to cooperate.

Another problem is recruitment. The design demands enlistment of nursing homes or low-income households, and the experimenter has to convince the required number of units to sign on. About half of them have to be told that they will not receive any new services but will be required to give periodic information. Locating participating units, explaining the conditions of the experiment, and convincing them to participate is no small task. Then there is the issue of when to tell participants that they might be in the control group and receive no service at all. Cook and Shadish (1994) provide a balanced discussion of the pluses and minuses of revealing the possibility of control group status at various points in the recruitment process. It is an important issue because if people (or organizations) refuse to participate because they know about the no-service possibility, the randomness of the assignment is compromised.

Another problem is being sure that the program is being implemented as planned. If, say, the state welfare agency is not delivering the job-search services it is supposed to be offering, i.e. the intervention is not on offer, the SE would be testing the effects of a phantom policy or of an unknown intervention of the agency's own devising. Results of the SE would be meaningless. From experience, researchers have learned the importance of monitoring the implementation of the intervention.

Probably the most basic design issue is implementing and maintaining randomization. Often researchers do not do the random assignment themselves. The operating agency selects participants for its programs and in the process is expected to assign participants to intervention and control groups according to the protocols prepared by the researchers. The actual assignment is "often carried out by a social worker, nurse, physician, or school district official" (Cook and Shadish 1994, 558). Sometimes these people misunderstand what they are expected to do, and sometimes they are tempted to use their professional judgement in assignment decisions. Researchers have learned that they must not only train agency staff but also maintain an oversight presence to ensure that assignment is indeed random.

Nor is that the end of the problem. What started as true randomized assignment may become undone as time goes on. In some cases the experiment does not enroll enough participants. Agency staff therefore may raid the control group to fill slots in the program. People labeled "controls" may in truth receive the intervention. Or, and this is inevitable, participants may drop out of the program and the study. That would be fine if they dropped out equally from intervention and control groups for similar reasons. However, it is usually more common for controls to drop out. They are not receiving services and they have less reason to persevere. For example, in the income maintenance experiments, higher drop-out rates were registered in the control group and in some of the experimental groups receiving smaller benefits than in the more generous benefit groups. The effect of differential drop-out is to compromise the equality of the groups. A selection bias is reintroduced.

In other cases, the control group may become contaminated by being inadvertently exposed to the intervention under study. Teachers receiving an experimental professional development course may share some of their new learnings with fellow teachers in their school, regardless of their official "control" status.

The list of complications goes on and on. As researchers have become more sophisticated over time and with experience, they have identified a host of further threats to the validity of SEs. Manski and Garfinkel (1992) suggest that some interventions might cause changes in norms and attitudes in the community, and the changed community attitudes would influence the success of the intervention. Heckman (1992) and Heckman and Smith (1995) have written that people who enlist in SEs may not be representative of people who would participate in full-scale programs. Moffitt (1992, 2004), too, has worried about "entry effects," the conditions of a full-scale program that would affect participants' behavior that do not show up in small-scale experiments.

Time

The worlds of research and policy do not work in tandem. Social experiments are time consuming, often taking many years to design, implement, and finally analyze and report results. The policy process meanwhile has moved forward and the results of a SE arrive in a new, changed policy environment. Research results may have little or no relevance in this changed policy world. For example, the health insurance experiment began at a time when the development of a national health care system was under active consideration, and the impact of cost sharing had real relevance. By the time the results of the experiment were known, the health care debate had petered out and national health care was no longer an imminent possibility. The relevance of the results was greatly diminished (Greenberg et al. 2003).

In the past it has often taken four or five years (or more) before experimental results were ready. The housing allowance experiment ran much longer. It studied the effect of giving housing allowances to low-income people not only on the families involved but also on the *supply* of housing. It had to go on long enough for landlords to increase the number of housing units available to recipients of allowances. The study ran (in two cities) for eleven years (Bradbury and Downs 1981).

On the other hand, some experiments are too short to produce convincing results. The nursing home incentive study ran for thirty months. Many nursing homes were evidently not willing to change their practices in response to the short-term monetary incentives. One of the sponsoring agency's reports states:

> To the participants [nursing homes] ... it may seem a very brief duration and there may be reluctance to make staffing, policy, and organizational changes which could affect their environment long after the experiment is concluded. (Greenberg et al. 2003, 107)

Yet even within that brief time period, the study was not able to catch the wave. By the time it was completed, political interest had moved away from incentives and toward regulation.

Foresight is not a particularly strong point of social science. Trying to figure out what policy issues will be lively at some future point is an exercise for a soothsayer. Knowing how rapidly the political canvas changes, knowing how volatile the complexion of government is these days with the country divided almost equally between Republicans and Democrats, knowing how policy windows open and shut as the economy changes, can we ever be confident that we are foreseeing an appropriate mix of interventions? Many people worry about issues of causation in experimentation. We worry about the clouded crystal ball. Fortunately or not, in recent years SEs have become more modest. As noted in the next paragraph, they are making do with available data, and they are taking less time to complete. But they are testing more modest initiatives.

Expense

Expense can limit the value that social experiments can provide to policy making. There is generally a direct relationship between the complexity of a research design and its cost. The more policy alternatives, settings, or types of participants tested, the more expensive is the experiment likely to be. Thus, cost plays a direct role in limiting the relevance of the findings of social experiments to particular policy questions. Over time, social experiments appear to be becoming simpler and consequently cost less. Greenberg et al. (1999) suggest that this is due in part to the increased use of administrative databases rather than special surveys, an increase in the likelihood that organizations that would run the program are the ones involved in the social experiment (as opposed to developing new programs run by the research organization), simpler designs with fewer groups, and shorter tracking periods for participants.

Limits on How Much Can be Tested

It is a rare experiment that can test all the variations in a particular policy that may be relevant to the question under study. Thus, the findings of social experiments are limited only to specific alternatives tested. SEs take place in a limited number of sites with a particular set of participants, and the findings may not generalize to other settings or participants. The time horizon is often truncated (although not in the health insurance experiment). Only a few social experiments can assess trade-offs among components of the intervention. Almost none are large enough to examine differences among multiple subgroups of the client population (the income maintenance experiments are an exception). Few examine the behavior of the staff implementing the program and so have little to say about practices that are associated with better or worse outcomes. Costs of the intervention are not always carefully calculated (for example, in the nursing home reimbursement experiment, officials were unable to separate costs of running the program from costs of the study (Greenberg and Shroder 1997)).

A distinction can be made between "black box" experiments, which test one or a few treatments, and "response surface" experiments that test a wide range of treatments (Greenberg et al. 2003; Burtless 1995). Examples of the latter are the income maintenance experiments of the 1960s and 1970s in which income guarantees and tax rates were varied across the treatment groups and the health insurance experiment in which cost sharing was varied across the groups. Greenberg et al. (2003) conclude that if the particular intervention that is being tested is still on the policy agenda when the experiment is concluded, the black box experiment would be fine. However, that is almost never the case. The advantage of the "response surface" experiment is that the design allows for the estimation of elasticities over a range of treatment options and its results can be used in later simulation models well into the future.

Small Effects

Social experiments almost never produce slam-dunk findings. If a proposed intervention were so obviously superior, there would probably be little reason to experiment. Most policy proposals are uncertain. The results of experimentation are often marginal. There are small gains in certain circumstances with some subpopulations. Interpretation becomes critical.

Because experimentation is such a difficult craft, the results are not always authoritative. Decisions about the course of the experiment have to be made all along the way. Compromises are made, sometimes in response to crises in the environment, sometimes to fit within a budget, sometimes to suit the skills of the available staff, sometimes to meet deadlines, sometimes in an attempt to answer new questions that emerge in the course of the study. Other researchers will critique the findings. They may reanalyze the data. They will come up with new models that they claim better account for the patterns in the data. The experiment can get captured by the research experts and become fodder for struggles for dominance.

Feasibility of Random Assignment for Organizational/Community Interventions

Some innovative policy ideas involve intervening in neighborhoods or systems or states. Rather than giving service to individuals one at a time, the proposed policy is designed to change the practices and culture of a larger entity. Examples include: changing the attitude of welfare offices so that staff priority is to place the client in a job; changing the practices in a neighborhood so that families, restaurants, and law enforcement agencies actively work to prevent youngsters from drinking alcohol; and changing the culture of a school system so that teachers and administrators actively welcome parents to participate in their child's education. To test ideas like these in an SE requires study not so much of individuals as of the units that are being altered—welfare offices, neighborhoods, or school systems. The interest is the behavior of the collectivity.

The obvious solution is to randomize the unit. A certain number of school systems or neighborhoods might be assigned randomly to the intervention or to a control group. However, as the size of the unit increases (say, to counties or states), fewer units can be studied. It is extraordinarily difficult and expensive to study a large number of neighborhoods or counties, and few studies have managed to go beyond ten or twelve. However, with only a limited number of cases, the laws of probability do not necessarily work. Any differences observed between the intervention group and the control group may be the result of chance. There are too few cases to even out the lumps of chance. Therefore, randomization of large units is a partial solution at best. Here is an issue where research innovations are needed and are currently being developed.

Another reason for the objection to random assignment is that a city is not a city is not a city, nor are neighborhoods interchangeable, or health systems or schools. Each of them has a history. Each has a set of established traditions. Each has a culture that has developed over generations. Each has attracted particular kinds of civic organizations and program staff and residents. Harlem is not the South Side of Chicago, which is not Watts. P S 241 in Brooklyn is not the same as the Condon School in Boston (Towne and Hilton 2004). Even if a researcher were randomly to assign neighborhoods, they wouldn't be totally comparable, and differences observed at the end might be due not so much to the intervention as to the whole complex of prior history and culture. For example, an evaluation of a program to promote nutritious food products randomly assigned supermarkets in Washington and Baltimore. The intervention group of markets placed nutritious products in favorable shelf locations and distributed fliers about nutrition. The control group did nothing. The measure of success was the customers' purchase of nutritious foods. Results showed that there were more differences between the two cities than between the experimental and control groups.

Ethics

Ethical issues have dogged experimentation since its beginning. People have displayed considerable concern with withholding a social good from one group regardless of degree of need. Practitioners are often loath to allow services to be allotted on the basis of chance, without exercise of their own professional judgement. Beneficiaries of service object strongly to being placed in a no-service control group. A host of ethical issues (withholding services for those eligible, full disclosure of experimental procedures, right to refuse, harm to participants) may significantly limit the questions that social experiments can address.

The rebuttal is that no one really knows whether the service is a social "good" until it has been studied. Many experiments find that the intervention is no better than standard service—or even detrimental. Thus, the nursing home reimbursement experiment did not show positive effects from the reimbursement scheme. Bickman's study of intensive mental health service, which included all the professionally fashionable bells and whistles, showed that intensive service did not have better results than regular service (Bickman 1996).

Complexity of Interventions

Perhaps the most vivid argument against experiments is that they assume that interventions have a simplicity that can be captured in a treatment/no-treatment design. Many interventions are highly complex social interactions, and simple cause-and-effect patterns may not be easily detected. The "program" is often implemented differently by staff, and the desired outcomes are social processes that cannot be readily measured by simple metrics. Studying the effects of psychotherapy, for example, poses all manner of problems because of the inherently personal ways in which therapists work and clients respond. No matter what label one affixes to the "brand" of psychotherapy, or how assiduously one tries to train therapists to use the same procedures, critics argue that quantitative randomized studies cannot yield sensible results.

Similarly, educators often say that interactions within a classroom, such as the introduction of a new teaching method, cannot be studied appropriately by quantitative randomized techniques. The assumption that all teachers trained in the new teaching method will implement it consistently, and that children in all classrooms will react in similar ways, represents a fundamental misunderstanding of the variability of teaching and learning. The rejoinder is that despite the variability, which certainly introduces more error of measurement, large samples should show the extent to which mean scores (of social functioning, of math achievement, of attendance) differ across populations exposed and unexposed to the intervention. In Cook's (2001) words: "It is not an argument against random assignment to claim that some schools are chaotic, the implementation of a reform is usually highly variable, and that treatments are not completely faithful to their underlying theories." There is enough consistency in human behavior, experimentalists claim, to allow an experiment to reach valuable conclusions about whether an innovation is worth adopting.

7. Conclusions

We started this chapter with a description of three distinctive traits of SEs: research in the field, conducted through random assignment of samples of prospective beneficiaries to intervention and control conditions, in order to test the probable success of a policy intervention. The first two characteristics are increasingly accepted as viable and necessary. Research in the field has now become mainstream practice. Randomized studies have received considerable support not only from the research community (although some researchers, particularly in the field of education, have lodged vigorous dissents) but also in Congress. For example, the education legislation that Congress passed in 2002 gives preference to evaluation studies with randomized designs. It is the third feature that may no longer be as firmly established: the prospective test of alternative policies.

SE came into prominence in the late 1960s at a time of turbulent policy change. It was part of the climate of innovation and radical reform that was sweeping the country. In the late 1980s and 1990s, as interest in fundamental change lessened, the fortunes of experimentation also shifted. Experiments continued to be done, more of them in fact, but fewer resources were devoted to them. The emphasis changed from major innovations to marginal improvements in existing programs. In Burtless's words, they were "narrower in focus, less ambitious, and less likely to yield major scholarly contributions" (1995, 63). Now, at a time of budget deficits and fiscal stringency in the USA and elsewhere, the likelihood of new domestic initiatives seems low. It is not a time when large new ideas will be tested, at least with government funds. The trend is to test minor modifications, preferably cost-saving modifications, and shifts of activity to the private sector. If you were considering investment in large-scale SEs, our advice would be: hold off. The product is a sound one, with high potential, but the time is not now—at least in the USA. But hang in. Some version of SEs will have their day.

We also began our story with an outline of three themes: the complexity of the policy world, the technical complexity of the research world, and the alignment or misalignment between experimental findings and policy questions. Overall, SEs have showed the possibilities and the limits of affecting policy through social science research. They have contributed considerable new knowledge. Some of their findings have infiltrated the policy arena and are part of policy-speak (Anderson 2003; Weiss 1999). Influentials in Congress, federal agencies, international organizations, interest groups, and the media learn to be conversant with experimental findings in order to take an informed part in policy conversation.

On the other hand, there are no examples of an SE that led directly to policy change. Results of the health insurance experiments were so late and so unfocused on actual legislative proposals that they were pretty much ignored—except by economists, who have used them to model new proposals. The nursing home reimbursement experiment results also arrived late, after the zing had gone out of the incentive idea. Almost nobody was still interested in incentives for nursing homes; the action was in the area of regulation. While widely published, the income maintenance experiments led to little concrete change in policy. The welfare-to-work experiments seemed to have policy consequences. The MDRC study provided support for mandatory work-first requirements and demonstrated the ability of states to design and manage their own welfare programs. All three of these program design aspects ultimately ended up in the Family Support Act of 1998. Nevertheless as we have seen, the experiment merely reinforced what policy makers were planning to do on other grounds.

Because policy making is such a complicated business, with so many players pursuing such divergent interests, it is overly optimistic to expect research information to carry the day. Even the high-quality information supplied by SEs cannot overwhelm all the other forces on the scene. And as we have seen, the timing of SEs is often off. The policy agenda moves on, while the SE is still studying last year's proposals.

Yet, totting up advantages and disadvantages, we come out in favor of further experimentation. The world is in dire need of greater understanding of the consequences of government action. Social experimentation cannot fully satisfy the needs for knowledge about policy outcomes, partly because of the intrinsic nature of social science research and partly because of the limitations imposed by the conditions under which it is done. Still it makes headway. Anything that advances rationality in the messy world of policy is worth supporting. Not venerated or kowtowed to, but cheered on.

But we also need to moderate our expectations of the contributions that SE can make. The notion of basing policy strictly on experimental evidence is wrong-headed. SE doesn't tell everything that a polity needs to know about a pending policy option.