Social Experimentation: Evaluating Public Programs with Experimental Methods

PART 7

Social Experimentation and the Policy Process

Larry L. Orr

March 6, 1998

OFFICE OF THE ASSISTANT SECRETARY
FOR PLANNING AND EVALUATION

Preface

This paper is part of a series on the design and implementation of social experiments. These papers are intended to provide a relatively non-technical synthesis of the fundamental principles of the evaluation of public programs using experimental methods, for both those who design and conduct social experiments and those who use the results of experimental studies. A complete listing of the papers in this series is provided below..

The author is indebted to Wendell Primus and David Ellwood, former Deputy Assistant Secretary and former Assistant Secretary for Planning and Evaluation, U.S. Department of Health and Human Services, respectively, for the financial support needed to produce these papers. I also wish to thank Howard Rolston, Steve Bell, and the participants in a series of seminars for the staff of the Office of the Assistant Secretary for Planning and Evaluation for their insightful comments and suggestions on earlier drafts of these papers. Erik Beecroft and Lu Nguyen provided computational assistance in developing some of the exhibits.

PAPERS IN THE SOCIAL EXPERIMENTATION SERIES

Part 1: Background and Rationale
Part 2: Basic Concepts and Principles
Part 3: Alternative Random Assignment Models
Part 4: Sample Design
Part 5: Implementation and Data Collection
Part 6: Analysis
Part 7: Social Experimentation and the Policy Process

The author is Chief Economist, Abt Associates, Inc. These papers were written while he was a visiting scholar in the Office of the Assistant Secretary for Planning and Evaluation, on leave from Abt.

Social Experimentation and the Policy Process

As a society, we have been guilty of what can be fairly termed policy corruption. In pursuit of bold visions, we have launched one bold scheme after another without anything like responsible evidence... The problem is not the visions. Americans across the political spectrum want to improve education, reduce violence, eliminate substance abuse, strengthen families, restore traditional values, and increase opportunity for achieving the Dream. The problem is that we know little more now than in the 60s about how, on a large scale, to achieve these shared objectives. And the reason is a continuing surrender to ignorance. Major public-policy initiatives are routinely advanced, but rarely do we organize to evaluate what works.

Richard Darman[1]
Director, U.S. Office of Management and Budget, 1989-93

This is the seventh in a series of papers on the design, implementation, and analysis of social experiments. In this paper, we consider the relationship of experimental evaluations to the policy process, in an effort to provide useful guidance to those who initiate, design, and interpret the results of social experiments. We begin by discussing the ways in which experimental evidence can inform the policy debate—and some of the ways in which it can be misused in that debate. We then consider the factors that increase, or decrease, the likelihood that experimental evidence will influence policy. Finally, we discuss ways in which experiments can be used more systematically in the policy process.

The Use of Experimental Evidence in the Policy Process

Most policy decisions are made on the basis of very little reliable information about the likely effects of the proposed policy. Indeed, those effects are often not even central to the policy debate—precisely because reliable evidence is lacking. In public discussions and legislative debates, and even in the internal deliberations of the executive branch, it is often simply *assumed* that proposed programs will achieve their stated objectives; the political debate then revolves around whether those objectives are worth the cost of the program. Even when the effectiveness of the program becomes an issue, anecdotes and empty rhetoric often pass for evidence.

In this environment, properly designed and executed social experiments can provide a unique, and extremely important, input: clear, convincing, and valid evidence of the impacts of the program on the outcomes it was designed to affect. In some cases, such evidence is sufficient to make the crucial difference in whether a policy is adopted or rejected, or a program continued or discontinued. As described in the first paper in this series, the Perry Preschool Project, the Manhattan Bail Bond Experiment, the Work-Welfare Experiments, and the National JTPA Study have all had clear, direct impacts on the adoption or continuation of specific policies or (in the case of JTPA) major funding changes for an ongoing program.

It is important to recognize, however, that the development of social policy is better understood as a *process*, rather than as a sequence of discrete "policy decisions". There is seldom a single point at which all the evidence and arguments relevant to a proposed policy are marshaled and an up or down decision made. Rather, a large number of actors in the policy process, both government officials and private citizens, engage in an ongoing dialogue in which policy is shaped incrementally, often as a compromise among the various political factions involved. These actors are, in turn, influenced by a large number of information sources and interested parties, both inside and outside government.

Given the multiplicity of actors, decision points, and information sources in the process, the policy effects of experimental evidence are usually indirect. In many cases, experimental results *affect prevailing attitudes and*

[1] Darman (1996).

opinions in a policy area. For example, although the Health Insurance Experiment did not lead directly to any specific policy changes, its finding that cost-sharing leads to reduced use of medical care without any discernible effects on health status was an important factor in the acceptance of cost-sharing as a cost-containment strategy, both in public programs and, perhaps more importantly, in private insurance plans.

Even when the experimental results do not clearly indicate whether a particular policy should be adopted or not, they can be very useful in *clarifying the trade-offs facing policy makers*. For example, the labor supply parameters estimated in the income maintenance experiments were incorporated into simulation models that were used to estimate the costs and behavioral effects of a number of welfare reform proposals in the 1970s. Similarly, the medical care demand elasticities derived from the Health Insurance Experiment in the early 1980s are still being used to predict the utilization effects and costs of health insurance policy proposals.[2] While such behavioral and cost estimates seldom constitute compelling evidence for or against a particular policy, they improve the information base on which policy decisions are made.

Experimental findings can illuminate policy trade-offs even when the outcomes of interest cannot be measured in monetary terms. The evaluation of the AFDC Homemaker-Home Health Aide Demonstrations, for example, found that the provision of home care to elderly and disabled clients did not result in the hoped-for reductions in hospital and nursing home costs. It did, however, improve the clients' mental functioning and sense of well-being.[3] Thus, policy makers were faced with the decision whether these nonmonetary benefits were worth the cost of the home care services. In this case, while the experiment could not decisively determine whether the program was socially worthwhile, it was able to lay out the relevant costs and benefits more clearly and accurately than any other available form of evidence.

Often the trade-offs delineated by the experimental evidence are between different subgroups of society. An evaluation of youth conservation and service corps, for example, found that the overall monetary benefits of the programs exceeded their costs.[4] When benefits and costs were calculated separately for program participants and the rest of society, however, net monetary benefits to participants were found to be positive, while the rest of society bore net monetary costs. Again in this case, the experiment was not able to determine conclusively whether the program was a worthwhile social investment, but it was able to quantify clearly the distributional trade-off involved in the policy decision.

It is important to recognize that, as these examples suggest, the measure of an experiment's social utility is *not* whether the program being tested is enacted or not—or, indeed, whether any specific change in policy results directly from the experiment. Rather, the measure of an experiment's social value is *whether it improves the information on which policy decisions are based*. An experiment that convinces policy makers *not* to adopt a new program that has net social costs is just as valuable as one that convinces them to enact a program with positive social benefits of the same magnitude.[5] Even if the experiment does no more than to confirm the preconceived views of one side in the policy debate, it performs the very useful function of strengthening the evidence that can be used by that side in arguing for (or against) the policy; this will improve the odds that the ultimate decision will be more beneficial to society.

Some potential misuses of experimental results

The power and credibility of experimental evidence can also be misused in policy debates. One way in which this can occur is through selective use of experimental results. An agency might, for example, attempt to suppress studies that do not support its preconceived policy positions. Fortunately, most social experiments are large and visible enough that it is hard to bury their findings entirely.

A more subtle (and more common) selective use of experimental results occurs when the sponsoring agency publicizes only those experimental findings that support its policy positions. Thus, for example, an agency might disseminate a summary of the research findings that highlights positive impacts, but neglects to mention significant adverse effects of the policy tested. An even more insidious selective use of findings sometimes occurs when the experiment estimates impacts on a large number of outcomes and finds effects that are significantly different from zero for only a small proportion of the

[2] See, for example, Rivlin et al. (1994), Jensen and Marlock (1994), and Ozanne (1996).

[3] See Orr and Visher (1987).

[4] See Jastrzab et al. (1998).

[5] For a formal model of the value of a social experiment, see Burtless and Orr (1986).

outcomes analyzed. As discussed in a previous paper in this series, there is a high risk that these are false positive results—estimated impacts that are significantly different from zero by chance alone. Focusing on this subset of results may therefore seriously mislead policymakers.

Another common misuse of experimental (and other) research results is to apply them to a policy or target population that is substantially different from those on which the results are based. Some mismatch in this regard is virtually unavoidable—given the lags involved in planning, implementing, and analyzing an experiment, research can seldom anticipate the exact policy or target population that will be of interest when the results become available. This means that the correspondence between the experimental intervention and population and the policy and population to whom the results are to be applied must be carefully assessed and appropriate caution used in interpreting the policy implications of the results. In some cases, the correspondence will be so weak that it is better simply not to try to apply the experimental results. In cases where the intervention tested is sufficiently similar to the policy under consideration, but the population on which it was tested is not, it may be possible to simulate the effects on the relevant population through reweighting of the experimental sample or simulation modeling that incorporates the effects found in the experiment.

Even when the experimental results can be taken to be a valid representation of the expected response to the policy of interest, it is important to remember that the fact that the intervention has an impact that is significantly different from zero does not guarantee that the policy is socially beneficial. Only a careful, comprehensive benefit-cost analysis can determine whether the program impacts are sufficient to justify their cost.

The misuse of research results is a natural outgrowth of the nature of the policy process. That process is basically an adversarial one, in which individuals routinely use new information selectively to support preconceived positions rather than objectively weighing the pro's and con's of every action. Analysts can do several things to counter the effects of this ideology-driven environment.

First, they can encourage policymakers to distinguish between their ultimate objectives and the means to those objectives. All too often, policymakers believe that concern for any particular population or problem is synonymous with support for all existing or proposed programs intended to benefit that population or address that problem. It is important for them to realize that the program is a means to an end, and that if the program is ineffective, it may be worse then useless--like wearing a copper bracelet as a cure for cancer, it may divert attention and resources away from finding a truly effective solution to the problem. It is in the best interest of both the taxpayers who fund the program and the intended beneficiaries to obtain the best possible evidence with which to assess objectively whether the intervention is truly effective--and to scrap ineffective interventions. While this may seem like a truism, ideologically based support for, or opposition to, specific policies is perhaps the greatest obstacle to the effective use of research in the policy process.

A second way that analysts can combat the misuse of research in the policy process is to make sure that, in their own reports, the results are laid out as clearly and completely as possible, with all the appropriate caveats, and that they are widely disseminated. This will allow partisans on both side of the issue to use whatever support the findings provide for their own positions and to challenge the other side's use of the results. In effect, this strategy seeks to use the adversarial nature of the policy process itself to police the misuse of research results. While this approach has some obvious shortcomings—e.g., subtle caveats about statistical inference are quickly lost in a world of sound bites—in the long run, it provides the best hope of raising the informational content of the policy debate and, therefore, leading to better policy decisions.

Factors That Affect the Likelihood Experimental Results Will Influence Policy

The flip side of the danger that research results will be misused in the policy process is the danger that they will not be used at all. We hasten to add that by "used" we do not necessarily mean "lead to the adoption of a new program." In practice, it is rare that a single study is decisive with respect to any given policy; fortunately, research can play an extremely valuable role even if it does no more than raise the quality of the policy debate. To achieve even this more limited objective, though, the research must be taken into account by the actors in the policy process as they fashion policy.

On the basis of a review of the literature on research utilization, Greenberg and Mandell (1991) suggest that five characteristics of an evaluation will condition the

degree to which it will influence policy—its credibility, timeliness, communication and visibility, generalizability, and relevance.[6] In addition, they argue that the utilization of the results will be strongly affected by the policy environment in which they are considered. We consider each of these factors in turn.

Credibility

It may seem obvious that the more credible a piece of research is, the more likely it is to influence decision-makers. But the credibility of research results depends on a complex interplay of factors, not all of which have to do with the scientific quality of the research. This is true in part because policymakers are generally not qualified to judge the scientific quality of research; therefore, they must rely on indirect indicators of the reliability of research evidence—e.g., the reputation of the researchers, whether the results are generally accepted within the research community, and whether the results are internally consistent and consistent with the users' own preconceptions and other evidence at their disposal. The complexity of the results is also an important factor; political actors tend to be suspicious of evidence that requires a complicated explanation.[7] Finally, the willingness of policymakers to give credence to any given set of results will be strongly conditioned by whether doing so would threaten that individual's political self-interest or established policy positions.

Experiments fare well on many, but not all, of these criteria. Perhaps most importantly, the fact that there is a consensus within the research community that experimental designs produce unbiased impact estimates tends to lead to broad acceptance of the results of experimental studies. While large, highly visible evaluations like the income maintenance experiments or the National JTPA Study naturally attract a good deal of scrutiny and some criticism in the research community, their results have been much less controversial than those of nonexperimental evaluations in these areas, such as the CETA studies discussed in the first paper in this series. The results of smaller experiments, such as the evaluations of state welfare-to-work programs, have generally been noncontroversial within the research community.

Experimental results should also fare well in terms of simplicity. The basic experimental method is quite straightforward—comparison of randomly assigned treatment and control groups is a concept that even lay persons can readily grasp. Nevertheless, analysis of multiple treatments, outcomes, and/or population subgroups can give rise to an imposing array of impact estimates; unless skillfully communicated, such results can give at least the impression of complexity. We discuss communication of experimental results in more detail below.

Similarly, experiments should rank high on internal consistency. Properly designed experiments yield results that are in fact internally consistent. However, the presence of sampling error can sometimes create the appearance of inconsistency among results if careful attention is not given to correct statistical interpretation. For example, an evaluation of a recent employment and training demonstration found that the intervention increased employment rates by a statistically significant 20 percent, but that the impact on earnings was not significantly different from zero. If one makes the common mistake of interpreting estimated impacts that are not significantly different from zero as if they *are* zero and estimated impacts that are significantly different from zero as if they are *exactly* equal to the point estimate, these two results are hard to reconcile. But if one takes sampling error into account, it becomes clear that there is substantial overlap between the confidence intervals around the two estimates. It could well be that the demonstration did have the same percentage effect on employment and earnings, but the demonstration sample size was simply not large enough to detect the impact on a high variance outcome like earnings as significantly different from zero. Alternatively, the impact on earnings may have been smaller than the impact on employment, although not necessarily zero

A different source of apparent inconsistency among experimental results arises when multiple tests of the same or similar interventions are conducted by different researchers. This was, for example, the case with the income maintenance experiments of the late 1960s and early 1970s. Although these four projects were popularly viewed simply as tests of the negative income tax, they in fact tested a number of very different interventions

[6] See the cited article for extensive relevant references to the literature on research utilization, as well as case studies of utilization of two important sets of experiments, the income maintenance experiments and the work-welfare experiments. The following section of this chapter owes much to Greenberg and Mandell.

[7] Quite aside from the credibility of complex evidence, it is also true that results that tell a simple "story" are more easily deployed in the give-and-take of policy debates than those that require detailed explanation in order to be understood.

(not all of which were negative income taxes) with very different populations, used different outcome measures, and presented their results in very different ways.[8] Not surprisingly, even the professional research community had difficulty sorting out the results.

Researchers can exercise a good deal of control over their peers' perceptions of the quality of the experiment and its simplicity and internal consistency by following sound methodological practice in its design and interpretation, and by presenting the results as clearly and simply as possible. The threat to the credibility of the results over which they have little control is the possibility that the results will conflict with policymakers' preconceptions, policy positions, or self-interest. And given the adversarial nature of the policy process, there will almost always be some subset of decision-makers for whom the results are unexpected or unwelcome.

Timeliness

Research results can only influence policy if they are available at the time policy actions are being considered. A traditional view of the policy process is that discrete policy actions are considered and taken (or rejected) within narrow "policy windows" defined by political events.[9] If this view is correct, then social experiments are at a decided disadvantage with regard to timeliness because of their long life spans. As noted in earlier papers in this series, a typical experiment takes at least 3-5 years to complete—and can take as long as 10 years from initiation to final report. Clearly, if an experiment is begun when interest in a particular policy issue is high, by the time it is completed the policy window will have long since closed. But attempting to predict which policy issues will be of interest 10 years in the future is a hazardous business; if experiments must anticipate policy windows that far in advance, most will miss the mark and end up being useless.[10]

Fortunately, in reality policy is made in a much more flexible, continuous way than the "policy window" model would suggest. Most policies and programs evolve over a number of years. For example, "welfare reform" has been a subject of concern in both the federal executive branch and Congress almost continuously for over 30 years. While the specific policies and programs proposed and enacted have varied widely over that period, many of the underlying behavioral issues (most notably, how to equip and/or motivate welfare recipients to work) have been remarkably constant over that period. Much the same could be said of policy with respect to the federal role in health insurance or job training. In such policy areas, experiments focused on fundamental behavioral issues that are central to the effects of policy are highly likely to be relevant no matter when their results become available. Thus, for example, while the income maintenance experiments were conceived as part of the Johnson's administration's Great Society and implemented during the Nixon administration, the results were actually used in the formulation of President Carter's welfare reform proposals.

It must also be borne in mind that the results of experimental research can have a relatively long shelf life. As noted earlier in this paper, some of the behavioral parameters estimated from the income maintenance and health insurance experiments in the 1970s are still being used in policy simulations today. Thus, even if the "policy window" is not open at the time the experimental results are released, those results are added to the inventory of knowledge about a particular policy area and are available to be drawn upon the next time the policy window is open. Of course, to be useful in subsequent policy rounds, the experiment must address important, fundamental behavioral issues or generic policy approaches that are of continuing interest.. A test of a single idiosyncratic policy package is likely to be obsolete by the time it is completed—if only because the champions of that particular approach are likely to have left the government.

Experiments with particularly compelling results may create their own policy windows. That is, policy issues that would not have otherwise come up for consideration may be thrust onto the agenda because of the results of an experiment or set of experiments. The UI Self-Employment Demonstrations are an example of this phenomenon. Conducted by the research office of the U.S. Unemployment Insurance (UI) Service as part of an effort to find ways to facilitate the reemployment of UI claimants, the demonstrations showed that providing training and financial assistance to claimants to help them start their own businesses was cost-effective from both

---

[8] At the time the experiments were being implemented, a prescient memo from the White House to the Secretary of Health, Education, and Welfare warned of a "cacophony of conflicting results" if too many different income maintenance experiments were launched. The memo was signed by President Nixon, but almost certainly written by Daniel Patrick Moynihan, then an advisor to Nixon.

[9] See, for example, Kingdon (1984).

[10] As one Congressional staffer who played a key role in the development of the Family Support Act of 1988 put it: "Finding out whether the [work-welfare] demonstrations worked better than what we had before was important. Having the data in time to help shape and promote our legislative efforts was nothing short of amazing" (Baum, 1991).

the claimant's perspective and the government's. These findings led directly to a legislative proposal to allow states to establish such programs. This proposal was enacted into law in 1993; to date, ten states have adopted enabling legislation for self-employment allowances and seven have implemented programs under this provision.[11]

Finally, one type of experiment for which timing is much less an issue is the evaluation of ongoing programs. Because ongoing programs receive legislative scrutiny every year as part of the annual appropriations process, their policy window is virtually always open. Thus, the results will be relevant for policy whenever they become available unless the program has been eliminated or substantially changed while the experiment was in progress. Thus, for example, the final results of the National JTPA Study, which was launched in 1986, did not become available until 1994, but had an almost immediate impact on program policy via the appropriations process.

Communication and visibility

Research results can find their way into the policy process in any of a wide variety of ways—or not at all. Whether and how they are communicated will strongly condition the extent to which they are used by policymakers.

In this regard, it is important to note that "policymakers" can include a wide range of actors in both the executive and legislative branches at several different levels of government. Their level of understanding of social science research can vary equally widely--from those with essentially no familiarity (and often little patience) with technical material to those with advanced degrees in social science disciplines. Communicating with such a heterogeneous audience is difficult, both in terms of finding channels to reach them effectively and in terms of articulating the message in appropriate terms.

Sometimes the channels of communication that matter are relatively straightforward, as in the example described above of the UI Self-Employment Demonstrations. One of those demonstrations was Congressionally mandated at the initiative of a congressman with an interest in the policy. Once the results of the experimental tests became available, he sponsored a legislative proposal to enable states to adopt the intervention as an ongoing program. Because the proposal was backed by solid evidence that it would not cost the taxpayers money and was otherwise noncontroversial, the strong backing of a single congressman was sufficient to secure its enactment as Federal law.

Usually the lines of communication between researchers and policymakers are more indirect. In many policy areas there are established "issue networks" that link researchers, program managers, and policymakers.[12] This is true, for example, in the areas of welfare, employment and training, and youth programs. Members of such networks communicate through such diverse vehicles as professional meetings and conferences, committees and working groups formed to address specific issues, contractual relationships between private research organizations and government agencies, and personal contact, as well as through more formal mechanisms such as legislative hearings and dissemination of written documents. Professional organizations and public interest groups, such as the Association for Public Policy Analysis and Management, National Governors' Association, and the National Association for Welfare Research and Statistics, play important roles in these networks. Often research results are widely known within the network long before they are formally published as reports or journal articles. Although top-level policymakers seldom participate directly in such networks, their staffs frequently do; this provides one of the most important channels through which research results flow into the policy process.

The size and long life span of most experiments give them some advantage in gaining visibility within the policy community. In many cases, experiments have been widely discussed within the relevant issue network long before their results are available.

Finally, experimental results are much more likely to be influential if they have an advocate/interpreter who promotes them in the policy community[13]. It is not sufficient simply to publish a report of the findings and expect policymakers to act upon them. Someone has to bring the results to the appropriate policymakers' attention, explain how they relate to policies under consideration, and respond to questions and criticisms from both the research and policy communities. Such advocacy usually involves repeated (and repetitive!) presentations

[11] For a detailed description of the use of these experimental results for policy purposes, see Orr et al. (1994).

[12] For a detailed discussion of this concept, see Heclo (1978).

[13] I use the terms "advocate" and "promote" in the neutral sense of vigorously bringing the results to the attention of the research and policy communities, not in the partisan sense of using them to advocate particular policies. In practice, however, it is sometimes difficult to distinguish the two.

in the many forums of the relevant policy network; it may also involve direct communication with high level government officials or their staffs.

Perhaps the best example in recent years of such advocacy on behalf of research results is the efforts of Judith Gueron in advancing the results of the work-welfare experiments of the early 1980s. Her numerous presentations at conferences and research meetings led to working directly with the chairman and staff of the Senate Subcommittee on Social Security and Family Policy in the drafting of the Family Support Act of 1988, the most important welfare reform legislation of the 1980s. The evidence from those experiments is generally credited with playing a key role in the passage of the key component of that Act, the Job Opportunities and Basic Skills (JOBS) program.[14] Certainly, those results were well known by participants in the legislative process—one observer-participant counted forty separate references to the studies in the public hearings on the Family Support Act.[15] While this level of visibility is rare for social science research, it illustrates the impact that experimental research can have if aggressively promoted.

Generalizability

Experiments test specific policies applied to specific populations in specific geographic areas. The utility of the results therefore depends crucially on how closely similar the specific experimental intervention, target population, and locale are to the policy context within which the results are to be applied. As noted above, the long life span of the typical experiment means that experimenters will seldom be able to anticipate exactly the intervention or program population that will be of interest to policymakers when the results become available several years later.

If the *experimental sample* overlaps the population that is of interest for policy, mismatches in composition can sometimes be addressed by analyzing subgroups of the experimental sample or reweighting the sample to match the policy population.[16] This might be the case, for example, if the results of a particular experimental intervention are available for the overall AFDC caseload in a state, but policy interest focuses on only those women who have been on the rolls for more than two years, or where the composition of the caseload has changed since the experiment was conducted. In cases where the experimental sample and the policy population do not overlap (e.g., where the experiment was conducted in a different state), it will be a judgment call whether the experimental results will provide more accurate guidance than the available evidence for a more closely similar population (e.g., nonexperimental studies of the AFDC population within the state).

In some cases, similar *ex post* adjustments can be used to address mismatches between the *experimental treatment* and the policy of interest. If the experimental treatment was defined as variations along a continuous policy dimension (e.g., tax rates or welfare benefit levels), to allow estimation of a behavioral response surface,[17] the response to policy parameters not directly tested in the experiment can be inferred by interpolation from the responses to experimental treatments. Thus, for example, the Health Insurance Experiment tested only four coinsurance rates (0, 25, 50, and 100 percent), but those four rates covered the policy-relevant range and responses to other rates can be inferred from the responses to those four by interpolation.

Unfortunately, only a minority of public policies can be characterized with continuous numerical parameters. Most experiments involve "black box" treatments—complex, multidimensional interventions whose overall impacts may reflect the effects of any or all of their component parts. To some extent, it is possible to decompose such treatments into their component parts at the design stage, through the use of factorial designs.[18] But there are both conceptual and practical limits to the number of separate program components whose effects can be separately identified and, in the end, one is still left with the fact that the components themselves are black boxes. For example, one might break a training program down into classroom training, on-the-job training, and job search assistance. But the impacts of, say, the classroom training component will be the result of a specific combination of curriculum, instructor's skills and background, physical facilities and equipment, program length and intensity, etc. If the proposed program to which the results of the experiment are to be applied differs in any of these dimensions, one cannot be sure that it would have the same effects as the experimental program.

[14] See Baum (1991) and Haskins (1991) for two views of the role of these projects in the legislative process.

[15] Haskins (1991).

[16] Both subsampling and reweighting involve some loss in precision—the former because of the loss of sample size, the latter because, for a given sample size, weighted estimates are less efficient than unweighted estimates. This is another instance of the tradeoff between precision and bias noted in earlier papers in this series.

[17] See the third paper in this series.

[18] See the third paper in this series.

It is tempting to conclude that, for these reasons, black box experiments will seldom be useful for policy and therefore should be avoided. That is almost certainly an overreaction. Especially in areas where there is a dearth of reliable research evidence, knowing with some certainty the effects of an intervention similar to the policy of interest may be extremely valuable, even if the two are not identical. And in cases where there *is* no preexisting "policy of interest," an experiment that demonstrates that a new intervention is cost-effective may generate substantial policy interest in the intervention that was tested.

One of the most difficult issues of generalizability facing experimenters is that of the *geographic representativeness* of the sample.[19] Because experiments generally require direct contact with the sample in order to administer the treatment and collect data, experimental samples are usually clustered in a small number of geographic locations, in order to keep costs manageable. In contrast, policy interest usually focuses on larger geographic areas, such as an entire state or the nation as a whole. A large number of environmental factors and participant characteristics that can potentially affect the impact of the experimental treatment vary across geographic areas. Unless the experimental sites were randomly selected from all possible sites in the larger universe, there is no guarantee that these factors, and therefore the experimental impact estimates, will be representative of the larger universe.

Experimental sites are usually not randomly selected, for several reasons.[20] Sites are frequently chosen in ways intended to ensure cooperation with the implementation requirements of the experiment. For example, researchers sometimes issue invitations to participate to large numbers of organizations of the type that will be required for the experiment; the location of those organizations that volunteer then determines the experimental sites. Even where researchers have attempted to select a national probability sample of sites, they have not always been successful, because of the refusal of organizations in many of the selected sites to accept random assignment of applicants to a no-service control group.[21] And in many cases, funding constraints limit the experiment to such a small number of sites that even if they were randomly selected, their representativeness would be questionable.

Researchers sometimes deliberately forgo random selection of sites in favor of studying sites with "interesting" interventions or "best practices." Unfortunately, while that approach may yield information about the effectiveness of those particular approaches, in the end the question of how generalizable those practices are may prevent policymakers from acting on the experimental results.

In contrast, nonexperimental studies are often conducted on nationally representative data bases collected for other purposes, such as the Current Population Survey or the decennial census. Until researchers find ways to select more generalizable experimental sites, then, nonexperimental analyses will often have the advantage with regard to this criterion. This does not necessarily mean that the nonexperimental results are more reliable. But it does mean that policymakers are sometimes faced with a choice between internally valid experimental evidence that is of questionable external validity (i.e., unbiased estimates for an unrepresentative experimental population) and externally valid nonexperimental results that may not be internally valid (i.e., potentially biased estimates for a representative sample).

Relevance

Research is obviously more likely to influence policy decisions if it is seen as relevant to the issues that are central to the policy debate. In this regard, experimental research has the advantage that it focuses on behavioral responses to interventions that are within the control of policymakers, in contrast to, say, research that seeks to understand social interactions and behavior without linking them to policy.

Not all behavioral responses to policy interventions are central to the decision to adopt or retain the intervention, however. Some programs are justified on the grounds that they further certain social principles or values, almost without regard to their effects on behavior. Thus, for example, research demonstrating that the Social Security program has adverse effects on the labor supply of older workers is unlikely to convince policymakers to eliminate Social Security. Unless there is some program feature (e.g., Social Security's treatment of earned income) that

[19] See the fourth paper in this series.

[20] Two experiments that were successfully implemented in a nationally representative set of sites are the evaluation of the Food Stamp Employment and Training Program (Puma et al., 1990), which had 53 randomly selected sites, and the Job Corps evaluation (Burghardt et al., 1997), which was implemented in all 111 program sites nationwide.

[21] See the discussion in the fourth paper in this series of the experience of the National JTPA Study in this regard.

can be adjusted to mitigate the effects found by the research, such findings are likely to be ignored. In designing experiments, therefore, it is essential to identify the specific policy decision that might be influenced by the experimental results and to assess the importance in that decision of the behavioral responses being measured.

The perceived relevance of social research will also depend on its timeliness and generalizability, which were discussed earlier in this section.

The policy environment

The likelihood that experimental results (or any other research) will influence policy also depends on the policy environment into which they are injected. As noted at the outset, research is only one of many factors that enter into policymakers' deliberations. If research is to influence the outcome of those deliberations, the other factors must be sufficiently inconclusive or offsetting for research evidence to tip the balance one way or the other. Moreover, policy decisions to which the research is relevant must either be "on the table" or the research must be sufficiently persuasive to convince policymakers to take up those issues. The latter happens only infrequently.

In a widely accepted view of the policy process, Weiss (1983) summarizes the influences on policy as "the interplay of ideology, interests, and information." In Weiss's model, ideology is driven primarily by principles and values, and "interest" is defined primarily in terms of the policymaker's self-interest, not the social interest. Because policymakers' ideologies and the interests they represent are relatively impervious to empirical information, and research is only one of many sources of information on which they rely, it might seem that research is destined to play only a very marginal role in policy. Indeed, one of the major implications that Wiess draws from her model is that the greater the consistency of ideology, interest, and other sources of information, the less influence research is likely to have in the policy process.

A more optimistic view of the process is that ideology and interests set the *objectives* of policy, but not the methods by which those objectives can be achieved. The latter is an empirical issue on which research can shed light. Thus, even when ideology and political interest are agreed upon a particular objective, there is still room for experimental evidence to influence the means chosen to attain that objective.

An example of such a situation is the policy environment in which the results of the National JTPA Study were released. The federal government was running large deficits, which both the newly elected Republican Congress and the Clinton administration had sworn to eliminate. In this budget-cutting atmosphere, the response to the experimental results showing that JTPA had virtually no effect on the earnings of youths was to reduce the budget for that component of the program by roughly 80 percent. The budget for the adult component was left virtually intact, however, largely on the basis of the study's finding that the adult component was cost-effective. Thus, in this case the experiment was able to show policymakers how to achieve their objective of achieving budgetary savings with the least loss to society.

There are, of course, cases where means as well as ends are dictated by ideology and political interest. But at a minimum, the existence of credible, relevant experimental evidence, clearly and prominently presented, makes it more difficult to justify approaches that conflict with that evidence.

Using Social Experimentation More Systematically in the Policy Process

While one can cite scattered success stories, if experimental research is to play a major role in the policy process, it must become a more routine part of the way government agencies evaluate existing and proposed policies and programs. Although a few agencies have begun to develop systematic programs of experimental research, all too many of the experiments that have been conducted to date represent the isolated triumphs of a few persistent individuals over a system that is not attuned to the experimental method. Moreover, the level of resources currently devoted to evaluation overall is an order of magnitude too small to allow systematic examination of the many existing and proposed programs and policies.

In this section, we discuss ways in which experimentation could be used more systematically to enlighten the policy process. Some of the approaches we discuss have already been adopted by at least one federal agency; others have yet to be implemented. The overall approaches considered here are:

- Systematic evaluation of ongoing programs;
- Testing multiple approaches to the same policy objective;
- Replicating apparently successful interventions; and,
- Mandatory testing of new policies.

Systematic evaluation of ongoing programs

It is unfortunate but true that we have little hard evidence of the effects of most ongoing public programs. And although most government agencies (at least at the federal level) have research and evaluation budgets, few of them use those resources to *systematically* evaluate the impacts of each of their ongoing programs to determine whether they are meeting their objectives. Rather, research and evaluation activities tend to focus on collecting descriptive data and testing new policy prescriptions.

As noted elsewhere in this series of papers, the payoff to evaluating ongoing programs can be quite high. If a social program is not producing the benefits to participants that are its *raison d'etre*, its elimination can save the taxpayers many times the cost of the evaluation required to measure its effects. Moreover, elimination of a program that is not producing the intended benefits to its participants entails little or no loss to those participants. In fact, continuation of an ineffective program may well *harm* its intended beneficiaries, not only because it wastes their time and creates unfulfilled expectations, but also because if policymakers assume that the program's objectives are being achieved they will not initiate a search for more effective approaches. If, on the other hand, the program *is* effective, it is important to establish that fact, so that it will not be scaled back or eliminated on the basis of less reliable evidence.

The tendency not to evaluate ongoing programs, in the face of the fairly obvious benefits of doing so, is probably attributable to several factors. First, the top levels of government tend to be preoccupied with justifying new programs and policies, rather than re-examining existing ones. Thus, it is easier to obtain resources to test a new idea than to evaluate an ongoing program. Second, experimental evaluations require the exclusion of the control group from the program, which program staff often find more ethically problematic in ongoing programs than in special demonstrations.[22] Third, unlike tests of new policy proposals, evaluation of an ongoing program threatens an existing bureaucracy, whose wages constitute the "taxpayer savings" that would be realized if the program is found to be ineffective. Finally, as noted at the outset of this paper, many programs are justified on ideological or political grounds, and it is simply *assumed* that they have their intended effects.

A notable exception to this rule is the evaluation program of the Employment and Training Administration (ETA) of the U.S. Department of Labor. Beginning with the National JTPA Study, which started in 1986, ETA has systematically launched large-scale experimental evaluations of each of its major ongoing employment and training programs--JTPA, the Job Corps, and the Economic Dislocation and Worker Adjustment Assistance (EDWAA) program.

The Job Corps evaluation, which is underway as this is written, is particularly noteworthy because of several novel features designed to address the difficult problems encountered in evaluating ongoing programs.[23] First, the evaluation sample is a random subset of all eligible applicants to the Job Corps in the 48 contiguous states and the District of Columbia; thus, the results will be generalizable to the national program. This was possible because, unlike many federal programs, the Job Corps is administered directly by the federal government, rather than through grants to state and local governments. Thus, it was not necessary to obtain the voluntary agreement of local programs to participate in the study.

This not only allowed the researchers to draw a nationally representative sample, but also to spread the sample thinly across all 111 local programs, rather than concentrating it in a small number of sites, as most previous evaluations had done. This in turn permitted the second notable design feature of the Job Corps evaluation: only about 7 percent of all eligible applicants were assigned to the control group.[24] The fact that only a small number of controls were drawn from each local Job Corps program substantially reduced the impact of random assignment on local program recruitment requirements and operations, as well as diminishing the resistance of program staff to the implementation of random assignment.

[22] See the discussion of ethical issues in the first paper in this series.

[23] See Burghardt et al. (1997).

[24] The 6,000 control group members were excluded from the program for 36 months. Of the 75,000 eligible applicants assigned to the treatment group, only 9,400 were included in the research sample, to keep the costs of follow-up data collection manageable. Thus, the treatment-control ratio in the research sample was approximately 3:2.

A limitation shared by all prior experimental evaluations of ongoing programs, including the Job Corps study, is that they have been onetime efforts, providing a "snapshot" measure of program effectiveness at a single point in time. Because programs evolve and change over time-indeed the problems they are designed to address may change over time-a program that is cost-effective today may not be a few years from now. Therefore, for policy purposes, it would be highly desirable to have a more continuous measure of program effectiveness.[25]

A modified version of the Job Corps evaluation design could provide such a measure. Instead of drawing the sample at a single point in time, one could assign a small proportion of all eligible applicants to a control group *on an ongoing basis* and *continuously* collect follow-up data on the outcomes of all program and control group members. Such a design would allow estimation of impacts for each annual cohort of participants and, because it would be an ongoing system, would provide much longer follow-up than is typical of onetime studies. In addition, by pooling samples from consecutive years one could obtain much more precise impact estimates for the overall sample and/or samples large enough to yield reliable estimates of program impacts on small subgroups of participants. Because data would be collected and analyzed continuously, an ongoing evaluation system would probably also reduce the lag between the program period under study and the time that impact estimates become available. (One would, of course, still have to wait at least two years to obtain two years of follow-up data.) Finally, short-term impact estimates would be much more informative for policymakers because they could be compared with the short-term impacts of the program on earlier cohorts, for whom longer-term impact estimates are available.

Although continuous random assignment has never been implemented in an ongoing program,[26] it is certainly technically feasible from an administrative and implementation standpoint. Even in decentralized programs, random assignment could be conducted as part of the regular intake process using PC-based software. It would require only the political will to provide the necessary resources and make random assignment a program requirement.

Testing multiple approaches to the same policy objective

As noted earlier, most social experiments have been tests of new programs or policies. Unfortunately, the tendency has been to test only one approach (or class of approaches) at a time, rather than a range of alternative approaches to the same problem. As a result, if the tested approach turns out not to be cost-effective, policymakers are left with no useful guidance as to how to address the problem. Only by initiating a new test of a different approach, which will take years to complete, can they hope to obtain a workable policy prescription. With *seriatim* testing of individual programmatic approaches, it could take decades to identify an effective policy intervention. And even when a cost-effective approach is identified, there is no assurance that it is the *most* cost-effective approach.

A better strategy would be to test a range of alternative policy options simultaneously, in a single, integrated research project. This could drastically shorten the time required to identify an effective approach. Properly designed, it would also ensure that the alternatives tested are truly comparable, in terms of their participants and the local environment.[27]

In those cases where multiple interventions have been tested experimentally, the alternatives have usually not been fully comparable. For example, the National JTPA Study estimated impacts on participants' earnings for several different service strategies.[28] Participants were not randomly assigned to different service strategies, however; rather, program staff selected the strategy deemed most likely to be helpful to the participants.[29] Thus,

[25] Some programs (e.g., JTPA) have a continuous Aperformance measurement@ system based on the post-program outcomes of program participants. Because such systems lack control groups, however, they cannot measure the impact of the program on participant outcomes. At best, they measure the relative impacts of different local programs. And if outcomes in the absence of the program would vary across local programs, they are not even a reliable measure of relative impact.

[26] We are aware of only one serious proposal along these lines. In 1991, the Food and Nutrition Service (FNS) of the U.S. Department of Agriculture issued for public comment proposed regulations that would have allowed state Food Stamp Employment and Training programs the option of using random assignment to measure program performance on an ongoing basis, rather than an outcomes-based performance measurement system (Employment and Training Reporter, 1991). These regulations were never adopted.

[27] See the discussion of tests of multiple alternative approaches in the third paper in this series.

[28] See Orr et al. (1996).

[29] This design was dictated by the objective of measuring the impacts of the program as it existed in the field Bincluding staff discretion to assign participants to the services they deemed most appropriate. Given this objective, the rationale for estimating the impacts of different service strategies was to find out which parts of the program were working well and which were not, rather than to compare the service strategies to find out which were the most effective.

differences in impact across service strategies may have represented differences in participant characteristics, as well as differences in program effectiveness. Similarly, a large number of different interventions intended to help welfare recipients move into employment have been tested in the past 20 years, but almost always in different sites, so that differences in program effects are confounded with site differences.[30]

These examples reflect the difficulty of implementing random assignment to multiple interventions in the same site, especially within the context of an ongoing program. In the case of the JTPA study, program staff were unwilling to allow random assignment to replace their professional judgment in the assignment of JTPA applicants to alternative service strategies. In the evaluation of welfare-to-work programs, local welfare programs were generally unwilling to take on the administrative complexities involved in running two different welfare-to-work programs side by side.

These kinds of implementation problems are very real and must be taken into account in designing tests of alternative policy interventions.[31] The returns to overcoming such problems can be substantial, however. By systematically testing multiple alternative interventions in the same setting, policymakers can obtain much more reliable policy guidance, more quickly, than can be derived from a collection of single-intervention experiments.

Replicating apparently successful interventions

Occasionally, an intervention appears to be quite successful on the basis of evaluation results in a single site or trial. When the evidence of success is based on a random assignment evaluation, policymakers and researchers alike have tended to accept such results as definitive evidence of program effectiveness. This can lead to a rush to apply whatever program features were believed to be unique to that site on a larger scale. The problem with accepting such evidence at face value is that it may reflect nothing more than the unique local environment within which the test was run or sampling variability in the assignment of the experimental sample.

This is especially true when (as is often the case) the "successful" program is the one site in a multi-site experiment where positive impacts were found. As noted in the previous paper in this series, one would expect to find impacts that are significantly greater than zero at the .10 level by chance alone in one site out of ten. Add to this statistical risk of false positive results the fact that any given intervention is likely to be genuinely more effective in some local environments than in others, and it becomes clear that statistically significant impacts in one site are not necessarily replicable in other sites.

Even when the "successful" program is not part of a multi-site experiment, a kind of selection bias that tends to bring false positive results to the fore may be at work. If, as one prominent evaluator has suggested, careful evaluation will show most social interventions to be unsuccessful,[32] any program that shows significantly positive results is likely to receive a great deal of attention. But by the same token, a high proportion of these apparently successful programs are likely to among the one in ten trials whose statistically significant results reflect only sampling error. If, for example, only one out of 100 interventions tested were truly effective, over 90 percent of the tests with significantly positive results would be false positives! (I.e., in 100 trials, we would expect one true positive and ten false positives.)

It is also true that, in social programs, it is not always clear exactly what the intervention was. What was implemented in the field may be quite different from the program model specified by those who designed the test. Even those who operate the program in the field may describe the program quite differently than it is actually run. And process analysts employed by the evaluator to document program operations may focus on the wrong subset of the thousands of details that make up even the simplest program.

For all these reasons, it is hazardous to base policy on a single, small-scale test of a new idea, however successful it may appear to be. This is not to say, however, that such results should simply be ignored. Such interventions are, after all, more likely to be successful than totally untested ideas. But they should be subjected to further validation before being adopted as policy.

---

[30] A partial exception to this pattern is the evaluation of the Job Opportunities and Basic Skills (JOBS) program, in which welfare-to-work programs that focused on immediate job search and placement were contrasted with programs that emphasized long-term education and training In three sites, welfare recipients were randomly assigned to both types of program. In the remaining four sites, however, only a single program was implemented. See Hamilton et al. (1997).

[31] See the discussion of implementation issues in the fifth paper in this series.

[32] Peter Rossi's "Iron Law of Evaluation" states that "the probability that any given social program will prove to be effective is not high."

An instructive example of an intervention of this type is the "job club", or self-directed group job search, a technique for helping the unemployed find jobs that was first evaluated in the early 1970s. In a random assignment evaluation of the original job club program in a single site, post-program employment rates in excess of 90 percent were recorded for the treatment group, in contrast to 55-60 percent employment rates for controls at the same point in time.[33] When the same approach was applied to unemployed workers with labor market handicaps, the results were even more dramatic: employment rates were still 90-95 percent for the treatment group, in contrast to control rates of only 20-30 percent.[34] Over the 25 years since the original job club experiment, this approach has been experimentally tested in a wide variety of settings for a broad range of clients—it may in fact be the most extensively evaluated intervention in the history of social experimentation. And while the impacts of self-directed group job search are often found to be significantly positive, they have also frequently been nonexistent and have never been as dramatic as those of the first few studies that brought the technique to the attention of national policymakers. Far from the panacea that they initially appeared to be, job clubs have turned out to be just one more moderately effective tool in the employment and training service kit.

The follow-up studies of the job club approach were, in most cases, undertaken in the context of broader studies, not as conscious replications of the initial study. A more deliberate policy of testing the replicability of promising findings has been pursued by the Department of Labor (DOL) in the case of education and training interventions directed toward youths. This effort grew out of the National JTPA Study's finding that JTPA had essentially no impact on the earnings of youths. In response to this finding, DOL consulted a wide range of experts in employment and training and youth development in an attempt to identify approaches that might be more effective than traditional JTPA services.

Two promising approaches were identified. The Quantum Opportunities Program (QOP) is a high school mentoring program that had been found to have strong positive effects on a wide range of outcomes, including graduation rates and performance on standardized tests, among disadvantaged students in a small-scale random assignment study in four cities.[35] The Center for Employment Training (CET) is an employment and training service provider whose San Jose, California, center had been the only site out of 13 to have significantly positive impacts on the earnings of youths in an experimental test of intensive training programs for youths.[36] Mindful of the danger that the sites or participant populations involved in these studies may have been atypical, or that the results may have simply been false positives, DOL elected to replicate these programs in a larger number of sites and evaluate the replication programs with random assignment, before attempting to implement them on a broader scale. The CET program is being tested in seven sites, while QOP will be replicated in twelve.[37]

Mandatory testing of new policies

Another way to protect against the risk of ineffective interventions being adopted as ongoing programs is to require that new policies be evaluated experimentally before being implemented on a permanent basis. This is analogous to the Food and Drug Administration requirement that new drugs pass randomized clinical tests of effectiveness before they can be put on the market.

In the context of U.S. social programs, this approach is particularly appropriate in programs where state or local governments have a good deal of policymaking discretion, with federal funding and oversight. In such cases, new programmatic approaches are continually being implemented, usually with little or no evaluation. For example, over the past 20 years, states have adopted a large number of education, employment, and training programs and financial incentives to help welfare recipients become self-sufficient. Many of these new program components required federal waivers of the state's approved plan; in the late 1980s, the U.S. Department of Health and Human Services began to require that these "waiver projects" be rigorously evaluated, usually with random assignment. Not only has this requirement forced states to objectively assess policy changes that were often launched with great political fanfare and overblown promises, but over the years a large body of evidence has

[33] Azrin et al. (1975).

[34] Azrin et al. (1979).

[35] Hahn (1994). A fifth site was unable to implement the program and was dropped from the evaluation.

[36] See Cave et al. (1993).

[37] See Maxfield (1997) for a description of the background and design of the CET replication study.

accumulated with regard to a wide range of interventions.[38] While this evidence is not as systematic as one might wish, it nevertheless provides valuable guidance to states considering interventions that have been tried elsewhere.

[38] See Greenberg and Shroder (1998) for descriptions of these evaluations.

References

Azrin, N. H., T. Flores, and S. J. Kaplan. 1975. "Job-Finding Club: A Group-Assisted Program for Obtaining Employment," *Behavior Research and Therapy, 13*, pp.17-27.

Azrin, N. H., and Robert A. Philip. 1979. "The Job Club Method for the Job Handicapped: A Comparative Outcome Study," *Rehabilitation Counseling Bulletin, 23*, pp. 144-155.

Baum, Erica B. 1991. "When the Witch Doctors Agree: The Family Support Act and Social Science Research," *Journal of Policy Analysis and Management. 10, 4, 603-615.*

Burghardt, John, Sheena McConnell, Alicia Meckstroth, and Peter Schocet. 1997. "Implementing Random Assignment: Lessons from the National Job Corps Study." Princeton, N.J.: Mathematica Policy Research Inc.

Burtless, Gary, and Larry L. Orr. 1986. "Are Classical Experiments Needed for Manpower Policy?" *Journal of Human Resources 21* (Fall): 606-39.

Cave, George, Hans Bos, Fred Doolittle, and Cyril Toussaint. 1993. *JOBSTART: Final Report on a Program for School Dropouts*. NewYork: Manpower Demonstration Research Corporation.

Darman, Richard. 1996. "Riverboat Gambling With Government," *New York Times Magazine, pp. 116-117.* December 1.

Employment and Training Reporter. 1991. "Proposed Food Stamp Regs Released." September 18, p. 27.

Greenberg, David H., and Mark Shroder. 1997. *The Digest of Social Experiments*. Second edition. Washington, D.C.: Urban Institute Press.

Greenberg, David H., and Marvin B. Mandell. 1991. "Research Utilization in Policymaking: A Tale of Two Series (of Social Experiments)," *Journal of Policy Analysis and Management. 10, 4, 633-656.*

Hahn, Andrew. 1994. *Evaluation of the Quantum Opportunities Program (QOP): Did the Program Work?* Waltham, MA: Brandeis University, Center for Human Resources.

Hamilton, Gayle, Thomas Brock, Mary Farrell, Daniel Friedlander, Kristen Harknett. 1997. *Evaluating Two Welfare-to-Work Program Approaches: Two-year Findings on the Labor Force Attachment and Human Capital Development Programs in Three Sites*. Washington, D.C.: U.S. Department of Health and Human Services and U.S. Department of Education.

Haskins, Ron. 1991. "Congress Writes a Law: Research and Welfare Reform," *Journal of Policy Analysis and Management. 10, 4, 616-632.*

Heclo, Hugh. 1978. "Issue Networks and the Executive Establishment", in Anthony King (ed.), *The New American Political System*. Washington, D.C.: American Enterprise Institute.

Jastrzab, JoAnn, John Blomquist, Julie Masker, and Larry Orr. 1997. *Youth Corps: Promising Strategies for Young People and Their Communities.* Bethesda, MD: Abt Associates Inc.

Jensen, Gail A., and Robert J. Marlock. 1994. "Why Medical Savings Accounts Deserve Another Look," *Journal of American Health Policy, pp. 14-23.* March/April.

Kingdon, John W. 1984. *Agendas, Alternatives, and Public Policies*. Boston: Little, Brown.

Maxfield, Myles, Jr., and Allen L. Schirm. 1997. *The Quantum Opportunity Program Demonstration: Year 1 Report.* Washington, DC: Mathematica Policy Research, Inc.

Orr, Larry L., Howard S. Bloom, Stephen H. Bell, Winston Lin, George Cave, and Fred Doolittle. 1996. *Does Job Training for the Disadvantaged Work? Evidence from the National JTPA Study*. Washington, D.C.: Urban Institute Press.

Orr, Larry L., Stephen A. Wandner, David Lah, and Jacob M. Benus. 1994. *The Use of Evaluation Results in Employment and Training Policy: Two Case Studies*. Unpublished paper presented at the Annual Research Conference of the Association for Public Policy Analysis and Management.

Orr, Larry L., and Mary G. Visher. 1987. *AFDC Homemaker-Home Health Aide Demonstrations: Client Health and Related Outcomes*. Washington, D.C.: Abt Associates Inc.

Ozanne, Larry. 1996. "How Will Medical Savings Accounts Affect Medical Spending?" *Inquiry 33:225-236.*

Puma, Michael J., Nancy R. Burstein, Katie Merrill, and Gary Silverstein. 1990. *Evaluation of the Food Stamp Employment and Training Program.* Alexandria, VA: U.S. Department of Agriculture, Food and Nutrition Service, Office of Analysis and Evaluation.

Rivlin, Alice M., David M. Cutler, and Len M. Nichols. 1994. "Financing Estimation and Economic Effects," *Health Affairs 13:30-49.*

Weiss, Carol H. 1983. "Ideology, Interests, and Information: the Basis of Policy Positions," in D. Callahan and B. Jennings (eds.), *Ethics, Social Science, and Policy Analysis.* New York: Plenum Press.