The Politics of Random Assignment: Implementing Studies and Impacting Policy

Judith M. Gueron
Manpower Demonstration Research Corporation (MDRC)

As the only nonacademic presenting a paper at this conference, I see it as my charge to focus on the challenge of implementing random assignment in the field. I will not spend time arguing for the methodological strengths of social experiments or advocating for more such field trials. Others have done so eloquently.[1] But I will make my biases clear. For 25 years, I and many of my MDRC colleagues have fought to implement random assignment in diverse arenas and to show that this approach is feasible, ethical, uniquely convincing, and superior for answering certain questions. Our organization is widely credited with being one of the pioneers of this approach, and through its use producing results that are trusted across the political spectrum and that have made a powerful difference in social policy and research practice. So, I am a believer, but not, I hope, a blind one. I do not think that random assignment is a panacea or that it can address all the critical policy questions, or substitute for other types of analysis, or is always appropriate. But I do believe that it offers unique power in answering the "Does it make a difference?" question. With random assignment, you can know something with much greater certainty and, as a result, can more confidently separate fact from advocacy.

This paper focuses on implementing experiments. In laying out the ingredients of success, I argue that creative and flexible research design skills are essential, but that just as important are operational and political skills, applied both to marketing the experiment in the first place and to helping interpret and promote its findings down the line. Conducting a successful random assignment experiment in a complex, real-world context requires a continuing balancing of research ambition against operational realism. Lest this sound smug — because I am talking from an institutional track record of success in doing this — let me add that success remains an uphill battle. No one has ever welcomed random assignment. Moreover, this challenge has recently become more acute as research questions and programs become more complex, and the political and funding terrain more hostile.

My background theme is that this is a battle worth fighting. People who are active in public policy debates and who fund this type of research know the political and financial costs of evaluations that end in methodological disputes. Henry Aaron put this well in his influential book *Politics and the Professors*, which describes the relationship

[1] See, for example, the papers by Boruch, Snyder, and DeMoya, 1999, and Cook, 1999, prepared for this conference.

between scholarship and policy during the Great Society era and its aftermath. Pointing to the conservative effect on policymakers of disputes among experts, he asked: “What is an ordinary member of the tribe [that is, the public] to do when the witch doctors [the scientists and scholars] disagree?” He went further, arguing that such conflict not only paralyzes policy but also undercuts the “simple faiths” that often make action possible.[2]

Random assignment, because of its unique methodological strengths, can help avoid this kind of conflict — what Aaron called “self-canceling research.” But random assignment studies must be used judiciously and interpreted carefully to assure that they meet ethical norms and that their findings are correctly understood. It is also important that researchers not oversell this technique. Random assignment can answer the important “Does it make a difference?” and “For whom?” questions, but it must be combined with other approaches to get answers to the critical question of “Why?” and “Under what conditions?”

BACKGROUND

Over the past 25 years, MDRC has conducted 30 major random assignment studies, in more than 200 locations, involving close to 300,000 people. Projects have ranged from the first multisite test of a real-world (that is, not researcher-run) employment program operated by community organizations (the National Supported Work Demonstration), to the first projects that moved social experiments out of the relatively contained conditions of specially funded programs into mainstream welfare and job training offices (the Work Incentive [WIN] Research Laboratory Project and the Demonstration of State Work/Welfare Initiatives), to what may have been the first efforts to use large-scale experiments to decompose the “black box” of an operating welfare reform program and determine the effects of its different components (the Demonstration of State Work/Welfare Initiatives and the more recent National Evaluation of Welfare-to-Work Strategies).[3] We have integrated random assignment into large bureaucracies (welfare offices, job training centers, courtrooms, public schools, and community colleges) and smaller settings (community-based organizations). The studies have targeted different populations, occurred in greatly varied funding and political contexts, and involved denying people access to services viewed as benefits (for example, job training to volunteers) or excluding them from conditions seen as onerous (such as time limits on welfare). We have been called names and have been turned down more times than accepted, but have so far managed to ward off legal challenges and avoid any undermining of the random assignment process. Although our experience shows ways to succeed, it also points to the vulnerability of this type of research and thus the need for caution in its use.

[2] Aaron, 1978, pp. 158–159.

[3] See Hollister, Kemper, and Maynard, 1984; Gueron, 1991, 1997; Leiman, 1982; and Hamilton et al., 1997. The National Evaluation of Welfare-to-Work Strategies (NEWWS) is an ongoing study, formerly titled the Job Opportunities and Basic Skills (JOBS) Evaluation, that was conceived and funded by the U.S. Department of Health and Human Services.

Because of this experience, I was asked to address two topics: What are the preconditions for successfully implementing a random assignment experiment? What are the preconditions for having an impact on policy? As hinted at above, I will argue that thinking in terms of “preconditions” is the wrong concept. It is true that there is soil that is more or less fertile, and some that should be off-limits, but, to continue the metaphor, the key to success lies in how you till the soil and do the hard work of planting and harvesting. You have to understand the context and clear away potential land mines.

This paper presents lessons from social experiments testing employment and training, welfare reform, and social service programs and systems. It first discusses the challenges in implementing a random assignment study and then the strategies to promote success and some guidelines on how staff should behave in the field. It then turns to the attributes of a successful experiment and the future challenges to using this approach.

Throughout this paper, I will use a number of terms. Any evaluation must differentiate between the test program’s *outcomes* (for example, the number of people who get a job or graduate from school) and its *net impact* (the number who get a job or graduate who would not have done so without the program). The measure of net impact is the difference between what would have occurred anyway and what actually happened because of the program.

A random assignment study (also called a social experiment) uses a lottery-like process to allocate people to the two or more groups whose behaviors (outcomes) are subsequently compared to determine the program’s net impact. People in one group are enrolled in the test program, and the others are enrolled in a control group intended to show what would have happened in the absence of the program, that is, to provide a benchmark, or counterfactual, against which to assess the program’s accomplishments (to determine its value added). Less frequently, the experiment may be a differential impact study, wherein people are assigned to two or more test programs (or two programs and a control group), with the goal of determining both the net impact and the relative effectiveness of the two (or more) approaches.

It is the randomness of this process — producing a control group that provides a convincing and unbiased estimate of the counterfactual — that makes this approach so powerful. Other strategies to estimate net impacts face the challenge of identifying an alternative benchmark that can be defended as providing a reliable measure of what would have happened without the intervention.

Administrators often know and tout their program’s outcomes, but they rarely know the program’s net impacts. In addition to the perceived administrative and ethical burdens of implementing a random assignment study, one reason this approach is not always welcome is that outcomes tell a more positive story than impacts. As a result, the challenges in launching a random assignment study are not only explaining this difference between outcomes and impacts but also convincing administrators that they want to know about — and can sell their success based on — net impacts.

SUCCESS IN IMPLEMENTING A SOCIAL EXPERIMENT

If someone is unreservedly enthusiastic about participating in the study, he or she doesn't understand it. [MDRC field rep]

Successful implementation of a social experiment means overcoming a series of hurdles, including:

- addressing the right question,
- meeting ethical and legal standards,
- convincing people that there is no easier way to get the answers and that the findings are "good enough,"
- balancing research ambition against operational reality,
- implementing a truly random process and assuring that enough people actually get the test service,
- following enough people for an adequate length of time to detect policy-relevant impacts,
- collecting reliable data on an adequate number of outcomes so you don't miss the story,
- assuring that people get the right treatment and enforcing this over time.

In this section, I discuss each of these in turn, focusing on the burden the issue places on operating programs.

This paper's litany of challenges may sound unrelenting, leaving the reader wondering why any manager would want to be in such a study. The reasons unfold below, but key among these have been the opportunity to learn (from the study and other sites), the potential to contribute to national and state policy, pressure (from the federal government or state officials) to evaluate program achievements, special funding, and, critically, the fact that the burden on staff was much less than originally feared. These reasons may sound abstract, but they have been sufficient for many sites to participate in repeated random assignment studies, even when earlier findings were not positive.

Addressing the right question

The first challenge is to be sure that the evaluation addresses the most important questions. Is the key issue net impact, or feasibility, or replicability, or what explains success or failure, or cost-effectiveness? If it is net impact, is the question: (1) "Does the XYZ service achieve more than the services already available?" or (2) "Are services, such as XYZ, effective?" or (3) "Is one service more effective than another?" Once it is clear what question you want to answer, the next challenge is determining whether you can design and enforce a social experiment to address it. The answer may be "no."

This "Compared to what?" issue may sound simple, but we have found it to be the most profound. The tendency in program evaluations is to focus on the treatment being assessed: Make sure it is well implemented so that it gets a fair test. While this is critical, our experience suggests that it is as important to define the treatment for the control group, because it is the difference in experience that you are assessing.

The challenge arises from the fact that social programs do not occur in a laboratory, thus limiting the researchers' ability to structure both the test and the control environments. With adequate attention and realism, you can usually get the test treatment implemented, but, for legal, ethical, and practical reasons (see the next section), there are severe limits on how much you can structure the control environment. Specifically, you cannot exclude control group members from all services available in their community or school. This means that you can usually answer question 1 above; for example, "Does the test training program or school reform do better than the background of existing services (to the extent that they are normally available and used)?" If this is the right question, the evaluation will satisfy the policy audience. But if the policy question is "Are the services provided of value at all?" (question 2) and if people in the control group have access to some level of similar services, the evaluation will fall short.[4] The difficulty is that people often agree up front that question 1 is the right one, but then they interpret the findings as though they had answered question 2.

There is no simple formula for getting around this issue, but it helps if the program being assessed is new, scarce, or different enough from the background type and level of service (or the program with which it is being compared) so that there is likely to be a meaningful differential in service receipt. Otherwise, you risk spending a lot of energy and money to reach the unsurprising conclusion that the impact of no *additional* service is zero, despite the fact that the services themselves may be of great value.

Meeting ethical and legal standards

You want to do what to whom for how long? [Question from the field]

Since all random assignment studies affect who gets what services, it is imperative to take ethical and legal concerns seriously. Inadequate attention to these issues can provoke the cancellation of a particular study and can poison the environment for future work. Experience suggests that social experiments should:[5]

- not deny people access to services to which they are entitled,
- not reduce service levels,
- address important unanswered questions,

[4] Although this discussion focuses on this problem in social experiments, the same issue arises in many quasi-experimental, comparison group designs.

[5] See Boruch, 1997, Chapter 3, for a discussion of ethical and legal issues in random assignment experiments. Some of these are discussed in the following sections of this paper.

- include adequate procedures to inform program participants and assure data confidentiality,
- be used only if there is no less intrusive way to answer the questions adequately,
- have a high probability of producing results that will be used.

The first two points establish the threshold criteria. In some sense, randomly selecting who does and does not get into a program always involves the denial of service, but this issue is much less troubling when the study assesses a specially funded demonstration that provides enriched services that would not exist but for the research and where the number of applicants substantially exceeds the number of program slots. Under those circumstances, random assignment can be viewed as an objective way to allocate scarce opportunities. Since the control group retains eligibility for all other services in the community, the experiment increases services for one group without reducing services for controls. Thus, when funds are limited and there will be no reduction in the level of service, random assignment can be presented as an ethical way to allocate scarce program slots, which at the same time will provide important answers as to whether the service is of value.[6]

It is more difficult to use a lottery to control access to existing services (for example, the regular job training system or a new welfare reform program). For certain individuals, such an evaluation will almost certainly lead to the denial of services that they would have received in the absence of the research. The key ethical demand in this case is to assure that the study is conducted only in locations where there are more applicants (or potential applicants) than available slots, and where the study, therefore, will lead to no reduction in the aggregate number of people served but only a reallocation of services among eligible applicants.[7] Of particular importance is avoiding any procedures that would deny people access to a service to which they are legally entitled (such as Medicaid or high school).

Suspicions about the ethics of researchers run deep, and despite attention to ethical and legal issues, MDRC staff have confronted numerous crises and, occasionally, horrific epithets. In one random assignment welfare reform study, county staff rejected participation, calling our staff "Nazis." In another state, a legislator accused our staff — and the state welfare agency funding the study — of using tactics similar to those in the infamous Tuskeegee syphilis study, provoking extensive negative press (including a cartoon characterizing the state as an unethical scientist pulling the legs off spiders just to see what happens). To save that study, state agency and MDRC staff had to meet with individual state legislators to explain the treatment for people in both the program and the control groups. (Program group members were required to participate in welfare-to-work activities and were subject to sanctions for nonparticipation; control group members were

[6] Even when there was no research purpose, administrators have sometimes used a random assignment lottery as a fair way to ration scarce and valued program opportunities, for example, in special "magnet" schools or the subsidized summer jobs program for youth.

[7] For an extensive discussion of the challenges and their resolution in such a study, see Doolittle and Traeger, 1990, Chapters 3, 4, and 6.

subject to neither condition, but would continue to have access to all entitlements, that is, Food Stamps, welfare, and Medicaid.)[8] We also stated what would be learned through the study, that we did not know whether the test program would help or harm people, and that there were not adequate funds to provide the test program to all people on welfare in the state. This process culminated in a state legislative hearing and, ultimately, a positive vote to endorse the study and random assignment.

Another example comes from the ongoing NEWWS Evaluation, where, in three sites, welfare recipients were assigned to a control group or one of two different treatments: one that pushes rapid entry into the labor force and another that stresses gaining human capital (primarily via adult basic education courses) before getting a job. Site staff were concerned that this random process would route people to services that did not meet their needs. The researchers responded that we were, in fact, undertaking the study because it was not clear which services were best for which people, an argument that ultimately proved persuasive.

In designing a social experiment, it is critical to determine how to inform people about the study and decide whether people can refuse to participate (the process of informed consent), to develop grievance procedures for controls, to protect the confidentiality of all data, and to limit the number of people in the control group. Most of these issues are straightforward, but that of informed consent is not. Researchers routinely use elaborate informed consent procedures in studies of voluntary employment and training or welfare reform programs that offer something of perceived value. At intake, individuals are told about the program, the intake lottery, and the data collection and confidentiality procedures and are offered the choice of participating in the study or not. Researchers have followed a different path in structuring evaluations of mandatory welfare reform programs, where the mandate was imposed by Congress or the state, not the evaluation. In this case, the research could not give people a choice of opting out of the program's requirements, because they were the law. Arrangements were worked out, however, to excuse a randomly selected group of controls from the new mandate, which might involve both services and financial penalties, including time limits on welfare. People in the program and control groups were informed that they would be in the study, told of the grievance procedures, and given a choice about participating in any surveys. The logic here was that (1) the study itself did not impose risks beyond those of daily life[9] and (2) controls continued to receive current services and were excused from a more restrictive program.

Finally, most large-scale field studies — whether or not they use random assignment — are expensive and are burdensome for program staff and participants. Funds spent on research may trade off against funds spent on services. Before launching

[8] The fact that controls were excused from a mandatory program that could involve grant cuts (rather than being denied a clear benefit) helped in defending this against the argument of service denial.

[9] This was the case because, in the absence of the study, people in the experimental group could not refuse to be in the new program (since it was mandatory) and because some people would routinely be denied services (since funds were limited); thus controls were not unduly disadvantaged. Moreover, all people in the study would continue to receive all basic entitlements.

such a study, the researchers should be sure that there is a high probability of getting reliable findings, that there is no less intrusive and less expensive way to get equally reliable results, and that the study has a high probability of addressing important questions and of being used.

Convincing people that there is no easier way to get the answers and that the findings are "good enough"

Over the past 25 years, as random assignment has proved feasible and research ambitions have grown, there has been a ratcheting up of study demands, making implementation increasingly challenging. Because of the service denial issue noted above, it was easier to promote participation in a small-scale test involving specially created programs than a random assignment evaluation of a large-scale ongoing program, especially one using a complex multi-group random assignment design. The ambitious, large-scale experimental tests of the Job Training Partnership Act (JTPA) and Job Opportunities and Basic Skills Training (JOBS) programs proved extremely difficult to launch, and many locations refused to participate.[10] One factor that helped enormously in promoting random assignment was evidence that the research community — not just the researchers conducting the study — had endorsed this approach as the most reliable way to determine net impacts. Of particular value were the findings of two national panels — the National Academy of Science's review of youth program evaluations and a U.S. Department of Labor panel's assessment of job training studies — that random assignment was the most reliable approach to determining the net impact of employment and training initiatives.[11]

It takes courage for political appointees to favor independent studies that measure net impacts. Aside from the normal desire to control the story, the challenge comes from the fact that impacts are almost always smaller than outcomes. For example, a job training program may accurately claim that 50 percent of enrollees got jobs, only to have this deflated by an impact study showing that 45 percent of the control group also found work, meaning that the program actually produced only a modest 5 percentage point increase in employment. It is much easier to sell success based on the 50 percent than the 5 percent, and particularly bedeviling to state that your program produced a 5 percentage point gain when another one (spared the blessing of a quality impact study) continues to trumpet its 50 percent achievement. I remember well the poignant question of a welfare official whose program we were evaluating. The governor had sent her a press clipping, citing outcomes to praise Governor Dukakis's achievements in moving people off welfare in Massachusetts, with a handwritten note saying, "Get me the same kind of results." She asked how our study could help, or compete.

[10] See, for example, Doolittle and Traeger, 1990.

[11] See Betsey, Hollister, and Papageorgiou, 1985; and U.S. Department of Labor, 1985.

Balancing research ambition against operational reality

Large-scale field research projects are rare opportunities. It is tempting to get very ambitious and seek to answer many important questions. Addressing some questions (for example, collecting more data on local economic conditions) adds no new burden on the operating program or study participants; addressing others clearly interferes with regular program processes. The challenge is to make sure that the research demands are reasonable, so that the program is not compromised to the point where it does not provide a fair test of the correct policy question or that the site is discouraged from participating in the study. Key decisions that can intrude on program processes include the degree of standardization versus local flexibility in multisite experiments, the extent to which sites must not change their program practices for the duration of the study, the point at which random assignment takes place, the duration of random assignment (and of special policies to serve experimentals and exclude controls), the intrusiveness of data collection, whether staff (as well as participants) are randomly assigned, and the use of multiple random assignment groups to get inside the "black box" of the program and determine which features explain program impacts.[12]

We have found that it is possible to implement random assignment, including large-scale studies in operating welfare offices, in ways that are not unduly burdensome. Among other steps, this has meant streamlining the random assignment procedures so that they take about a minute per person assigned. (In one site, the random assignment process became so routine that the site continued it after the study ended, viewing it as the most efficient way to match the flow of people into the program with staff capacity.)

Implementing a truly random process and assuring that enough people actually get the test service

Program staff generally dislike random assignment. This is true in community-based programs, where, to do their jobs well, staff must believe that they are helping people toward better lives. It is also true in large agencies, where it is feared that random assignment will add another routine for already-overloaded staff. While all of our studies were of programs funded at levels where not everyone could be served (so that access had to be rationed)[13] — and usually assessed services of unproved value (which, in some cases, ultimately were shown to hurt participants) — program staff vastly preferred to use their own, often more arbitrary rationing strategies (for example, first-come/first-served, serving the more motivated or more employable people, allowing caseworker discretion,

[12] The argument for randomly assigning staff arises in studies that compare two programs operating in the same offices or schools, in which staff or teacher quality may be a major explanation of program effectiveness. For an example of such a study, see Goldman, 1981. See Gueron, 1984, pp. 295 ff., for a discussion of the pros and cons of standardization. See Hamilton et al., 1997, and Miller et al., 1997, for examples of welfare reform evaluations that changed the nature or order of program services for some participants as part of a differential impact study; and Doolittle and Traeger, 1990, for a description of how the National JTPA Study took a different approach.

[13] Arguably, this is not true for our studies of time-limited welfare, although even in those cases, there were often accompanying services that could not be extended to all who were eligible.

serving volunteers first, or limiting recruitment so that no one was actually rejected) than to use a random process whereby they had to personally turn away people whom they viewed as eligible.

Yet random assignment is an all-or-nothing process. It doesn't help to be a little bit random. Once the process is undercut, the study cannot recover. To implement the study successfully, it is critical to get administrative and line staff to buy into and own the process. Two factors are central to achieving this. The first was already noted: reducing the burden that random assignment places on staff. This is where skill and flexibility in experimental design come in. You need to lodge an experiment in the complex program intake process in a way that minimizes disruption and maximizes intellectual yield. To do this, you have to understand the intricacies of recruitment and enrollment, the size of the eligible pool, and the likely statistical power of a sample under any particular design. One way to reduce the pain for staff is to place random assignment early in the process, before people reach the program office — for example, by randomly assigning students using centralized Board of Education records and then telling school staff to recruit only among those assigned as potential participants. This helps on one goal — reducing the burden on program staff — but it hurts on another: The longer the route from the point of random assignment to actual enrollment in services, the lower the percentage of people assigned to the test program who actually receive the treatment. This may mean either that the study has to get unrealistically large (and expensive) in order to detect whether the program had a net impact or that the study may fail to detect program impacts, even if they actually occurred.

The second factor in convincing program staff to join a random assignment study is showing them that the study's success has real value for them or, ultimately, for the people they serve. Two examples demonstrate how this was done. In 1982, when we were trying to convince state welfare commissioners to participate in the first random assignment tests of state welfare reform initiatives, we argued that they would get answers to key questions they cared about, that they would be part of a network of states that would learn from each other and from the latest research findings, that the study could give them cover to avoid universal implementation of risky and untested policies, that they would get visibility for their state and have an impact on national policy, that they would get a partially subsidized study, that randomly excluding people from service was not unethical because they didn't have enough money to serve everyone anyway, and, finally, that this technique had actually been used in a few local welfare offices without triggering political suicide.[14] Ultimately, eight states joined the study, which involved the random assignment of about 40,000 people in 70 locations and, in fact, delivered the benefits for the state commissioners that had been advertised.[15]

A few years later, MDRC launched a study that used random assignment to assess an education and training program for high school dropouts. To do this, we needed to

[14] For a discussion of how this was done, see Gueron, 1985.

[15] For a discussion of the impact of this study — known as the Demonstration of State Work/Welfare Initiatives — on national policy, see Haskins, 1991, and Baum, 1991.

find local providers who offered these services and convince them to participate in the evaluation. One such program was the Center for Employment Training (CET) in San Jose. CET leadership were dedicated to improving the well-being of Chicano migrant workers; the staff felt a tremendous sense of mission. Turning away people at random was viewed as inconsistent with that mission, and managers felt that the decision to join such a study would have to be made by the program intake staff — the people who would actually have to confront potential participants. We met with these staff and told them what random assignment involved, why the results were uniquely reliable and believed, and how positive findings might convince the federal government to provide more money and opportunities for the disadvantaged youth they served, if not in San Jose, then elsewhere. They listened; they knew firsthand the climate of funding cuts; they asked for evidence that such studies had ever led to an increase in public funding; they sought details on how random assignment would work and what they could say to people in the control group. They agonized about the pain of turning away needy young people, and they talked about whether this would be justified if, as a result, other youth gained new opportunities. Then they asked us to leave the room, talked more, and voted. Shortly thereafter, we were ushered back in and told that random assignment had won. This was one of the most humbling experiences I have confronted in 25 years of similar research projects, and it left me with a sense of awesome responsibility to deliver the study and get the findings out. The happy ending is that the results for CET were positive,[16] prompting the U.S. Department of Labor to fund a 15-site expansion serving hundreds of disadvantaged youth.

But even after getting site agreement on the rules, researchers should not be complacent. It is critical to design the actual random assignment process so that it cannot be gamed by intake staff. In our case, this has meant that we either directly controlled the intake process (that is, intake staff called MDRC and were given a computer-generated intake code telling them what to do, and we could later check that this was indeed followed), or we worked with the staff to assure that the local computer system randomly created program statuses.[17]

Following enough people for an adequate length of time to detect policy-relevant impacts

In conducting a social experiment, it is important to assure from the start that the sample is large enough and that the study will follow people long enough to yield a reliable conclusion on whether the program did or did not work. A sample that is too small can lead the researchers to conclude that an effective program made no statistically significant difference; a follow-up period too short may miss impacts that emerge over time.[18]

[16] See Cave et al., 1993.

[17] For a discussion of these procedures, see Gueron, 1985.

[18] See Boruch, 1997.

This may sound easy, but estimating the needed sample size requires understanding factors ranging from the number of people in the community who are eligible and likely to be interested in the program, the recruitment strategy, rates and duration of participation by people in the program, what (if anything) the program staff offer controls, access to and participation by controls in other services, sample attrition (from the follow-up data), the temporal placement of random assignment, and the likely net impact and policy-relevant impact of the program. Some of these factors are research-based, but others require detailed negotiations with the program providers, and still others (for example, the flow of people or the cost of data collection) may be clear only after the project starts. The complexity of this interplay between sample size and program operations points to the advantage of retaining some flexibility in the research design and of continually reassessing the options as operational, research, and cost parameters become clear.[19]

The pattern of impacts over time can be key to conclusions on program success and cost-effectiveness.[20] While this may seem to be primarily a data and budget issue, it usually also involves very sensitive negotiations about the duration of services provided to the program group, the length of time that control group members must be prevented from enrolling in the test program, and the extent to which the program can provide any special support for controls.[21]

Collecting reliable data on an adequate number of outcomes so you don't miss the story

A social experiment begins with some hypotheses about likely program effects. Researchers have ideas about these (usually based on some model of how the program will work), as do program administrators, key political actors, advocates, and others. We have found that, to get the buy-in for a study that will protect it during the inevitable strains of multi-year implementation, it is important to bring a diverse group of local stakeholders together and solicit their thoughts on the key questions. If people own the questions — if they see the project as *their* study that addresses *their* questions — they are more likely to stay the course and help you get the answers.

At MDRC, we learned this lesson in our first project that embedded random assignment in an operating social service agency — the WIN Research Laboratory Project of the 1970s. In proposing a partnership between staff in welfare offices and researchers, Merwin Hans (the U.S. Department of Labor WIN administrator) argued that local staff had undermined past studies because they did not care about the studies' success. To combat this, in this project the program staff were the ones who developed

[19] See Gueron, 1984, p. 293; and Doolittle and Traeger, 1990, for a discussion of this sequential design process in the National Supported Work Demonstration and National JTPA Evaluation.

[20] For example, our findings that different welfare-to-work programs have different time paths of impacts and that some produce taxpayer savings large enough to offset program costs depended on having data tracking people for several years after enrollment in the programs. See Friedlander and Burtless, 1995; Riccio, Friedlander, and Freedman, 1994; Hamilton et al., 1997; and Gueron and Pauly, 1991.

[21] In many studies, there is strong site pressure to provide some services to controls.

the new approaches and then worked closely with researchers on the random assignment protocols and research questions. Because they cared deeply about answering the questions, they provided the data and cooperated fully with the random assignment procedures.[22]

Designing field studies involves balancing research ambition against budget constraints. There is usually good reason to address a wide range of questions: for example, did an employment program affect earnings, transfer payments, income, family formation, or children's success in school? In deciding whether all this is affordable, a key issue is which data will be used to track behavior over time. The earliest social experiments (the Negative Income Tax experiments) relied on special surveys and researcher-generated data to track outcomes. The data were expensive but covered a wide range of outcomes. One of the breakthroughs in the early 1980s was the use of existing computerized administrative databases to track behavior.[23] These were much less expensive (allowing for an enormous expansion in sample size and, thus, a reduction in the size of effects that could be detected), placed less burden on study participants, and did not have the same problem of sample attrition faced in surveys; but administrative databases covered a narrow range of outcomes and had other limitations.[24] Moreover, gaining access to these critical administrative data can often be difficult and sometimes impossible, as state agencies may see little advantage in cooperating with the study and must balance research needs against privacy concerns.

In our early welfare studies, we argued for the value of answering a few questions well — that is, tracking large samples using records data — even if this meant we could address only the most critical questions. This seemed appropriate for studies of relatively low-cost programs, where modest impacts were expected and we therefore needed very reliable estimates to find out whether the approach made a difference and whether it was cost-effective. However, where programs are more ambitious and can potentially affect a wide range of outcomes for participants and their families, there is a strong argument for combining records and survey data, or using only survey data, to address a broader group of questions.

Identifying the data source is important, but it is also critical to collect identical data on people in the program and control groups. Estimating net impact involves comparing the behavior of the two groups. While it is tempting to use rich data on the program participants (about whom you usually know a lot), the key is to use identical data for people in the two groups, so that data differences aren't mistaken for program

[22] See Leiman, 1982, and Goldman, 1981.

[23] Examples of computerized administrative data include welfare and Food Stamp payment records, unemployment insurance data (which track people's employment and earnings), and various types of school records. The low cost of these data allows large samples to be followed over long periods, providing both more refined estimates of the impacts for the full sample and, equally important, estimates for numerous subgroups.

[24] See Kornfeld and Bloom, 1999, for a discussion of the relative merits of administrative records and surveys. While records data are relatively inexpensive to process, the up-front cost and time needed to gain access to these data can be high.

effects. Further, in all stages of the study, researchers need to be vigilant about data quality and comprehensiveness (thereby minimizing sample attrition).

Assuring that people get the right treatment and enforcing this over time

Random assignment is the gateway to placement in the different study groups. But a process that starts out random may yield a useless study if it is not policed. This means that, for the duration of the study, members of each research group must be treated appropriately; that is, they must be offered or denied the correct services. This is relatively easy if the test program is simple and controlled by the researchers. It is much more difficult if the program provides multidimensional services or is ongoing and operated in many sites, or if there is a differential impact study in which two or more program treatments are provided by staff in the same office.

To assure appropriate treatments and reduce crossovers (that is, people from one study group receiving services appropriate for the other group), staff need clear procedures on how to handle people in the different groups, adequate training, reliable systems to track people's research status over time, and incentives to follow the procedures. You need to be sure, for example, that if people return to a program (at the same or another office), they are placed in the same research status and offered the intended services. Obviously, the longer the treatment and the control embargo, the more costly, burdensome, and politically difficult is the enforcement of such procedures.[25] All these challenges, moreover, are multiplied in a differential impact study, especially when the two or more treatments are implemented in the same program office or school. In that case, it is particularly difficult to assure that staff or teachers stick to the appropriate procedures and that the treatments don't blend together, undermining the service distinction.

STRATEGIES THAT PROMOTE SUCCESS

The above discussion suggests some threshold "preconditions" that should be met to conduct a random assignment study: not denying people access to services or benefits to which they are entitled; not having enough funds to provide the test services for all people eligible; no decrease in the overall level of service, but rather a reallocation among eligible people; and, for programs involving volunteers, a careful process of informed consent.

Even if these conditions are met, successfully enlisting sites in a random assignment study is an art. As a neophyte to social experiments in the 1970s, I had thought that, to overcome the obstacles, it was critical that researchers have sufficient funding and clout to induce and discipline compliance with the requirements of the evaluation.[26] This surely helps, but as operating funds subsequently became scarce even while social

[25] See Gueron, 1985, pp. 9–10, for a discussion of these issues.
[26] See Gueron, 1980, p. 93.

experiments flourished, we learned that other factors could substitute. As noted earlier, key points were convincing the agency that the study would:

- advance its mission,
- provide the most reliable answers to questions the agency cared about,
- satisfy political concerns (for example, provide a way to avoid immediate large-scale implementation of an untested approach),
- get national visibility for the local program and its staff,
- follow ethical procedures, including, where appropriate, informed consent, full explanations of procedures, and a grievance process,
- satisfy federal or state research requirements or open up opportunities for special funding.

This last point has been particularly important. Obviously, states and sites would be more likely to participate in random assignment studies if this participation was a condition of their ability to innovate or get funds. This was one of several factors that explain the unusually large number of reliable, random assignment evaluations of welfare reform and job training programs. Key among these were that such studies were shown to be feasible and uniquely convincing, that staff at MDRC and other research organizations promoted such studies, and that staff in both the U.S. Department of Health and Human Services (HHS) and the U.S. Department of Labor (DOL) favored this approach.[27] Early studies (for example, the National Supported Work Demonstration and the WIN Research Laboratory Project) showed that random assignment could be used in real-world employment programs and in welfare offices. In the job training field, this success prompted the two prestigious review panels cited above to conclude that random assignment was superior to alternative evaluation strategies, leading DOL staff to fund both a large number of demonstrations that provided special funding to sites that would participate in such a study as well as a large-scale random assignment evaluation of the nation’s job training system.[28]

In the welfare field, HHS staff similarly became convinced of the value of random assignment and the vulnerability of other approaches. HHS staff were assisted in translating this preference into action by the requirement that Congress had put into Section 1115 of the Social Security Act, which allowed states to waive provisions of the Aid to Families with Dependent Children (AFDC) law in order to test new welfare reform approaches, but only if they assessed these initiatives. Since the early 1980s, through Republican and Democratic administrations, HHS staff took this language seriously and required states to conduct rigorous net impact studies.[29] In some states, there was also legislative pressure for such studies. The 1996 welfare reform legislation — the Personal

[27] In particular, Howard Rolston at HHS and Raymond Uhalde at DOL remained vigilant in promoting high-quality, rigorous evaluations.

[28] See Betsey, Hollister, and Papageorgiou, 1985; U.S. Department of Labor, 1985; and U.S. Department of Labor, 1995.

[29] For summaries of these studies, see Gueron, 1997; Gueron and Pauly, 1991; Greenberg and Wiseman, 1992; Greenberg and Shroder, 1997; and Bloom, 1997.

Responsibility and Work Opportunity Reconciliation Act (PRWORA) — substituted block grants for the welfare entitlement and ended the waiver process and evaluation requirements. No large-scale welfare evaluation using random assignment has been started under the new law.[30]

Other key points included showing that the study would not:

- undermine the program's ability to meet operational performance measures,
- reduce the number of people they served,
- overly burden hard-pressed line staff,
- deny controls access to basic entitlements or otherwise violate state laws and regulations, or
- likely lead to a political and public relations disaster.[31]

Finally, a number of other factors can make it more difficult to promote participation in a random assignment study:

- political concerns; for high-profile issues like welfare reform, public officials may prefer to control the data (using what they know about program outcomes) rather than risk more modest results from a high-quality independent evaluation,
- the perceived value of the services denied controls and the clout of members of the control group or their families,
- the intrusiveness of the research design (including the duration of any special procedures and the extent of interference with normal operations),
- the difficulty of isolating controls from the program (for example, from its message or similar services), which can limit the questions addressed in the study.

LESSONS ON HOW TO BEHAVE IN THE FIELD

I have argued that discovering which factors will induce participation and negotiating the design of an experiment that is politically and ethically feasible involve a balance of research and political/operational skills. To make this artistry less abstract, the following pages present some very basic operating guidelines that three senior MDRC staff members (Fred Doolittle, Darlene Hasselbring, and Linda Traeger) prepared for their colleagues to use as a starting point for more refined discussions.[32] As is clear from the tone, these were directed at staff seeking to enlist sites in a particularly challenging random assignment study of an ongoing operating program. In many studies, the site recruitment task is simpler, and this level of promotion is not needed.

[30] However, between 1996 and mid-1999, when this paper was completed, a number of small-scale, one-state studies were started.

[31] For an example of how these factors worked to bring states into the 1980s welfare experiments, see Gueron, 1985.

[32] Doolittle, Hasselbring, and Traeger, 1990; also see Doolittle and Traeger, 1990.

General rules

1. **The right frame of mind is critical. Remember, you want them more than they want you.** Even if initially they are eager, eventually they will figure out how much is involved and realize they are doing you a service if they say "yes." **Don't say "no" to their suggestions unless they deal with a central element of the study (for example, no random assignment).** You may well need to come back later with a modified design (for example, a different intake procedure) when the pickings of sites look slim. **Remember to be friendly and not defensive.** They really cannot know for sure what they are getting into, and their saying "yes" will be much more likely if they think you are a reasonable person they can work with over time.

2. **Turn what is still uncertain into an advantage.** When they raise a question about an issue that is not yet sorted out, tell them they have raised an issue also of concern to you and they can be part of the process of figuring out how to address it.

3. **Make sure you understand their perspective.** As much as possible, try to "think like them" so you will understand their concerns.

4. **Never say that something about the research is too complex to get into.** This implies they are not smart enough to understand it. Work out ways to explain complicated things about random assignment using straightforward, very concrete examples rather than research terms.

5. **Be sensitive about the language and examples you use.** Occasionally you will run into someone who has a research background and wants to use the jargon, but normal people are often put off by terms that are everyday, short-hand expressions to researchers. For example, many people find the terms "experiment," "experimental," "control group," "service embargo," and even "random assignment" offensive. Use more familiar, longer ways of saying these, *even if they are less precise or even technically wrong.* Site staff often react negatively to discussions of how random assignment is often used in medical research, probably because they are only familiar with outrageous examples.

6. **If some issues are sure to come up (ethics, operational issues, site burden), raise them yourself.** This shows that you understand the implications of random assignment, have grappled with them yourself, and think they can be addressed.

7. **If pressed on an awkward issue about random assignment, do not give an evasive answer.** For example, if site people forcefully ask if you really mean they will have to deny services to those in the control group, say "yes." Then,

explain the reasons for the rule, and address the underlying concerns that led them to raise the question.

8. **If someone is unreservedly enthusiastic about the study, he or she doesn't understand it.** While it might sound nice to let them cruise along happily, if their continued support matters, you must make sure they understand what they are getting into.

9. **Make sure you highlight the benefits of participating.** Usually, the key one is site-specific findings. Don't mislead them or allow them to think they will get more than you can deliver. Often, they want a lot of "inside the black box" type results.

10. **Negative momentum can occur and must be countered.** If things start going bad in many sites, regroup and rethink the model and the arrangements you are offering before things get out of hand.

Learning about the program

1. **Ask as many people as possible how the program works.** Different perspectives are vital. You need to know things at a micro level that only local people can know.

2. **Don't rely too much on their estimate of participation rates.** Unless they have an extraordinary management information system, most program operators have never had a reason to ask the type of client-flow questions needed to decide the details of a random assignment design.

Developing the details of the model and closing the deal

1. **Operational issues are your problem, and you have to get them to buy into the study before they become their problem.** You know you have made progress when they start helping you figure out how to address the problems.

2. **Don't be surprised by the level of "detail" you will have to address.** Something that seems like a minor point to you from a research perspective may turn out to be a crucial operational barrier to putting the model in place. Try to learn the vocabulary about the "details" so they will realize you understand and take their issues seriously.

3. **Realize that in working out procedures you will be dealing with people representing very different perspectives.** Program directors worry about different things than managers or the line staff. Be sensitive to the differences in perspective, and realize that a good director may give the managers who represent the line staff a veto over participation if you cannot address their

concerns. Support by an outside Board or director removed from program operations is not enough, although it is a start and will open the door. Administrative managers must be on board.

4. **Protect the core of the study, and figure out what you can give on.** Do not lose people over something not central. Depending on the study, noncentral items might include: who controls lists of people referred for random assignment, exclusion of certain groups of people from random assignment, temporary changes in the random assignment ratio to assure an adequate flow of program participants, length of the service embargo for controls, limited services after random assignment for controls.

5. **Sometimes the best response to a question about how a procedure would work is to ask a question in response.** The goal is to develop procedures for the study that disrupt the program as little as possible. When they raise a tough operational issue, the starting point is what they normally would do if the study were not in place. So ask them, and then go from there. Often, this will suggest minor changes that everyone can live with.

6. **Realize that model development is an iterative process.** New issues come up over time that will need to be addressed. Expect a continued balancing between research preferences and operational constraints.

7. **Develop a memorandum of agreement both parties can live with.** Don't push or even allow a site to sign an agreement you think they cannot fulfill. A key factor to be realistic about is sample size. Don't set targets they cannot meet.

8. **Money can often fix some problems, but don't get into a position where it looks as though you are trying to bribe them into betraying their ethics.** Operational issues relating to staffing can often be helped by financial support. Serious ethical concerns cannot be addressed in this way.

Community relations

1. **No news is good news.** Imagine yourself as a reporter. Would you rather write about the human interest side of the study ("Poor used as human guinea pigs") or the abstract policy and research issues that motivate the study? You should expect most local news stories done before findings are available to be negative if the reporter understands what random assignment is.

2. **Make sure the site knows you will take the bullets for them.** Convince the site that they have a compatriot who will join the battle if things get rough.

3. **There are pros and cons of your initially playing a prominent role in explaining the study.** Ideally, it would be best if the site took the lead in

building support for the study, because it shows they understand and really do support it. However, usually they can be surprised by local opposition or are not as good as you in explaining the reasons for the study or its procedures. If there is doubt how a meeting will go, fight for a role without implying that the local people don't understand the study or know the local situation.

4. **Be available to brief agencies affected by the study and advocates, but don't expect them to be won over instantly.** It takes a long process to convince someone that this type of research is OK. Make sure site staff understand the pros and cons of outreach to other groups versus a low profile. Then let the site staff decide how to play this.

5. **Prepare a press kit, and leave it up to the sites what to do with it.** This should be viewed as a defensive rather than an offensive weapon, to be used if called for.

6. **Develop a thick skin, and do not get defensive when speaking with the press or community groups.** There is one exception: If your personal integrity is attacked, fight back. You are not a "Nazi."

7. **Never say something is too complex to discuss or refuse to acknowledge key issues as legitimate.** Ultimately, participation involves trust. Random assignment isn't business as usual, and site staff have to know you are leveling with them.

Training local staff on study procedures

1. **Taking the time to write a good manual, with examples, is time well spent.** A detailed manual describing the study rationale and the intricacies of program intake and random assignment, and providing scripts for site staff, will serve as a valuable training tool and future reference for site staff.

2. **Realize that the training may be the first time many have heard much about the study and that you must win them over.** At the beginning of training, explain the reason for the study and random assignment and your common concern about people in the study. Try to get the site directors to lay the groundwork for the study and to show up at the training to indicate their support.

Setting the right tone for study implementation

1. **Program managers should understand that it is better to tell you about issues early, before they get serious and can threaten the study.** Try to convince people that you might be a source of possible solutions, based on MDRC's past experience.

2. **Make sure they understand you will show as much flexibility as possible on procedures.** Sites that decide to participate sometimes come to view the initial procedures as holy writ. They may nearly kill themselves trying to follow them without realizing you might be able to make a change that won't matter to the research but that will make their lives much easier. They probably will have trouble distinguishing between rules central to the core of the study and those that can be played with at the margins.

FROM RESEARCH TO POLICY: LESSONS FROM MDRC'S EXPERIENCE[33]

The previous sections of this paper discuss the challenge of implementing a random assignment study and the field techniques that promote success. But the ultimate goal of policy research is to inform and affect public policy. MDRC's studies have been credited with having an unusual effect on public policy, particularly welfare policy.[34] Looking back primarily at our welfare studies, I draw the following lessons about running a successful social experiment.

Lesson 1: **Correctly diagnose the problem.** The life cycle of a major experiment or evaluation is often five or more years. To be successful, the study must be rooted in issues that matter — concerns that will outlive the tenure of an assistant secretary or a state commissioner and will still be of interest when the results are in — and about which there are important unanswered questions.

Lesson 2: Have a reasonable treatment. An experiment should test an approach that looks feasible operationally and politically — where, for example, it is likely that the relevant delivery systems will cooperate, that people will participate enough for the intervention to make a difference, and that the costs will not be so high as to rule out replication.

Lesson 3: Design a real-world test. The program should be tested fairly (if possible, after the program start-up period) and, if feasible, in multiple sites. It is uniquely powerful to be able to say that similar results emerged in Little Rock, San Diego, and Baltimore. Replicating success in diverse environments is highly convincing to Congress and state officials.[35]

Lesson 4: Address the key questions that people care about. Does the approach work? For whom? Under what conditions? Why? Can it be replicated? How do benefits compare with costs? It is important not only to get the hard numbers but also to build on the social experiment to address some of the qualitative concerns that underlie public attitudes or that explain which features of the program or its implementation account for success or failure.

[33] This section is based on a discussion in Gueron, 1997, pp. 88–91.

[34] See, for example, Baum, 1991; Haskins, 1991; Greenberg and Mandell, 1991; Szanton, 1991; and Wiseman, 1991.

[35] Erica Baum stresses this point in Baum, 1991.

Lesson 5: Have a reliable way to find out whether the program works. This is the unique strength of a social experiment. Policymakers flee from technical debates among experts. They do not want to take a stand and then find that the evidence has evaporated in the course of obscure debates about methodology. The key in large-scale projects is to answer a few questions well. Failure is not in learning that something does not work but in getting to the end of a large project and saying, "I don't know." The cost of the witch doctors' disagreeing is indeed paralysis which, ultimately, threatens to discredit social policy research.

The social experiments of the past 25 years have shown that it is possible to produce a database widely accepted by congressional staff, federal agencies, the Congressional Budget Office, the General Accounting Office, state agencies, and state legislatures. When MDRC started its welfare studies, there was a football-field-long range of uncertainty around the cost, impacts, and feasibility of welfare-to-work programs. Twenty-five years of work have shortened this field dramatically.

Random assignment alone does not assure success, however. As discussed earlier in this paper, you need large samples, adequate follow-up, high-quality data collection, and a way to isolate the control group from the spillover effects of the treatment. You also need to pay attention to ethical issues and site burden. Finally, rigor has its drawbacks. Peter Rossi once formulated several laws about policy research, one of which was: The better the study, the smaller the likely net impact.[36] High-quality policy research must continuously compete with the claims of greater success based on weaker evidence.

Lesson 6: Contextualize the results. To have an impact on policy, it is usually not enough to carry out a good project and report the lessons. You need to help the audience assess the relative value of the approach tested versus others. To do this, you should lodge the results of the experiment in the broader context of what is known about what works and what doesn't.

Lesson 7: Simplify. If an advanced degree is needed to understand the lessons, they are unlikely to reach policymakers. One of the beauties of random assignment is that anyone can understand what you did and what you learned. One strategy we used was to develop a standard way to present results and stick to it. This meant that people learned to read these studies and understand the results. As social experiments are becoming more complex — involving multiple treatment groups and multiple points of random assignment — they put this overwhelming advantage at risk.

Lesson 8: Actively disseminate your results. Design the project so that it will have intermediate products, and share results with federal and state officials, congressional staff and Congress, public interest groups, advocates, academics, and the press. At the same time, resist pressure to produce results so early that you risk later having to reverse your conclusions.

[36] Cited in Baum, 1991.

Lesson 9: Do not confuse dissemination with advocacy. The key to long-term successful communication is trust. If you overstate your findings or distort them to fit an agenda, people will know it and will reject what you have to say.

Lesson 10: Be honest about failures. Although many of our studies have produced positive findings, the results are often mixed and, at times, clearly negative. State officials and program administrators share the human fondness for good news. To their credit, however, most have sought to learn from disappointing results, which often prove as valuable as successful ones for shaping policy.

Lesson 11: You do not need dramatic results to have an impact on policy. Many people have said that the 1988 welfare reform law, the Family Support Act, was based and passed on the strength of research — and the research was about modest changes. When we have reliable results, it usually suggests that social programs (at least the relatively modest ones tested in this country) are not panaceas but that they nonetheless can make improvements. One of the lessons I draw from our experience is that modest changes have often been enough to make a program cost-effective and can also be enough to convince policymakers to act. However, while this was true in the mid-1980s, it was certainly not true in the mid-1990s. In the last round of federal welfare reform, modest improvements were often cast as failures.

Lesson 12: Get partners and buy-in from the beginning. In conceptualizing and launching a project, try to make the major delivery systems, public interest groups, and advocates claim a stake in it so that they will own the project and its lessons. If you can do that, you won't have to communicate your results forcefully; others will do it for you.

One reason our research has had an impact is the change in the scale, structure, and funding of social experiments that occurred in the 1980s. The Supported Work and Negative Income Tax experiments of the 1970s were relatively small-scale tests conducted outside the mainstream delivery systems (in laboratory-like or controlled environments) and supported with generous federal funds. This changed dramatically in 1981, with the virtual elimination of federal funds to operate field tests of new initiatives. Most social experiments that we have conducted since then have used the regular, mainstream delivery systems to operate the program. There has been very little special funding.

The clear downside of this new mode was a limit to the boldness of what could be tested. You had to build on what could be funded through the normal channels, which may partly explain the modest nature of the program impacts. The upside was the immediate state and/or local ownership, since you were by definition evaluating real-world state or local initiatives, not projects made in Washington or at a think tank. If you want to randomly assign 10,000 people in welfare or job training offices in a large urban area, state or county employees have to have a reason to cooperate. When you are relying on state welfare and unemployment insurance earnings records to track outcomes, people have to have a reason to give you these data. The reason we offered was that these were *their* studies, addressing *their* questions, and were usually conducted under state contracts. They owned the studies, they were paying some of the freight, and thus they had a commitment to making the research succeed. In the welfare case, their commitment was aided by the fact that such evaluations also could satisfy the Section 1115 research requirements imposed by HHS.

Through this process, we converted state and local welfare and job training demonstrations and programs into social experiments, involving the key institutions as partners from the beginning. For the major actors and funding streams, the relevance was clear from the outset. This buy-in was critical. This partnership also had a positive effect on the researchers, forcing us to pay attention to our audience and their questions. In this process, during the 1980s and 1990s, social experiments moved out of the laboratory and into welfare and job training offices. Studies no longer involved a thousand, but tens of thousands of people. You did not have to convince policymakers and program administrators that the findings were relevant; the tests were not the prelude to a large-scale test but instead told states directly what the major legislation was delivering.[37] Because of the studies' methodological rigor, the results were widely believed. But the limited funding narrowed both the outcomes that could be measured and the boldness of what was tested.

Five years ago, I might have argued that these 12 factors explained why these studies had such a large impact on state and federal welfare policy. But that was clearly not the case in 1996. In contrast to the 1988 Family Support Act, which drew heavily on the research record, block grants and time limits are very much a leap into the unknown. While not necessarily pleasant, it is always useful for researchers to remember that their work is only one ingredient in the policy process and that, when the stakes are high enough, politics usually trumps research.

FUTURE CHALLENGES

Over the past two decades, random assignment studies have been used to build a solid foundation of evidence about the effectiveness of welfare reform and job training programs. In the early 1970s, it was not known whether this approach could be used to test real-world operating programs. We now know that it can be, and that the results are convincing. Although participation in random assignment studies involves clear burdens, administrators and staff in many programs have found the overall experience worthwhile and, as a result, have often joined multiple studies.

Yet the climate for such evaluations, at least in the welfare and job training fields, has grown chillier. Several factors explain this. One is the growing complexity of the research questions. Twenty-five years ago, the evaluation questions were very basic — Do employment and training programs make a difference? For whom? — and so were the random assignment designs. Thus, in the first random assignment test of such a program — the National Supported Work Demonstration — special funds were provided to small

[37] See Greenberg and Mandell, 1991, and Baum, 1991.

community programs to implement a clearly defined treatment; volunteer applicants were randomly accepted in the program or placed in a control group that got no special services. The study worked; the answers were clear.

Subsequently, the research questions have become more complex — What works best? What duration and intensity of the "treatment" produces what results? Which elements of a program explain its success or failure? Consequently, the operational demands have also grown in complexity, at the very time when there has been a reduction in special program funding. Random assignment moved out of small community programs into regular welfare and job training offices; tests covered not only special new programs but regular, ongoing services; studies involved not just one test treatment and a control group but multiple tests and more than one point of random assignment.

The result has been a major increase both in what is learned and — even more quickly — in what people want to learn. Random assignment has greatly increased the reliability of estimates of program impacts, but we have not progressed at the same rate in linking this to our understanding of program implementation. For example, most impact studies show substantial variation across locations, but they are not able to explain the extent to which this results from factors such as local labor market conditions or different aspects of program implementation. This limits researchers' ability to generalize the findings to other locations and also to get inside the "black box" of the program. Differential impact studies (comparing several approaches) are a major breakthrough, but realistically they can isolate the effects of only a few aspects of the treatment, or can compare only a few multi-dimensional approaches. They cannot provide an experimental test of the many separate dimensions of the program model and its implementation; yet this is the concern of people increasingly interested in understanding why initiatives produce the results they do and what should be done differently. More work in this area is critical if we are to increase the potential of evaluations to feed into the design of more effective programs.

A further complexity arises from the interest and need to assess saturation initiatives, for example, the end of the basic welfare entitlement or the launching of a comprehensive community-wide initiative. Random assignment has proved feasible in some cases, but not in others.[38]

Finally, funders and consumers of research are concerned that random assignment social experiments are intrinsically conservative, because the control group receives services regularly available in the community. In some studies, particularly of voluntary programs, the actual service differential may not be large. The resulting finding of a modest net impact leaves unanswered the question of whether the services themselves (received in varying forms and intensity by people in both groups) have a more

[38] For example, MDRC randomly assigned public housing communities in the Jobs-Plus Demonstration (see Bloom, 1999, and Riccio, 1999) but took a different approach in evaluating the effects of the 1996 welfare reform law in urban areas (see Quint et al., 1999). Also see Connell et al., 1995.

substantial impact.[39] Yet that may be the question uppermost on people's minds. The fact that this dilemma is not unique to random assignment studies, but is inevitable in any evaluation involving a comparison group, has not reduced the frustration. But it does mean that there is a hunger for a methodological breakthrough that would allow people to measure "total" not "net" impacts.

Finally, in the welfare field, the 1996 law, with its combination of block grants and the end of the Section 1115 waiver process, dramatically changed the funding and incentive structure that supported random assignment studies in the past. While block grants create pressure on states to figure out what works, the politicization of the welfare debate pushes in the opposite direction.[40] At a time when the stakes have never been higher for figuring out how to move people from welfare to work and out of poverty, the outlook for large-scale random assignment tests is unclear.

POSTSCRIPT: THE IMPLICATIONS FOR EDUCATION

This paper summarizes the practical lessons from random assignment studies of welfare reform and employment and training programs, but it is part of a series of papers addressing random assignment in education. Without straying from my topic unduly, some of the lessons suggest clear challenges for school-based random assignment studies. These include:

- Control services will be even more extensive, which has implications for the question you can address and the likely magnitude of impact.
- Controls will often be served in the same schools as experimentals, increasing the risk of the treatment's spreading from experimental to control classrooms.
- Treatments may extend over many years, making it harder to assure that people get the services defined in the research protocols.
- Schools are dynamic institutions, with many simultaneous innovations that can affect all students in the study.
- The unit of random assignment may have to be the school or the classroom.
- Teachers may be a key dimension of the treatment, raising the issue of teacher random assignment.
- Parents may pressure school principals to circumvent random assignment.
- The multidimensional experimental and control treatments will be more difficult to define and standardize across locations.
- The decentralized funding structure will both reduce the pressure for evaluation and increase the difficulty in disseminating research results.
- The involvement of children will make the implementation of informed consent and related procedures more demanding.

[39] See, for example, the discussion of the New Chance Demonstration in Quint, Bos, and Polit, 1997.

[40] See Gueron, 1997.

While this list is long, this paper points to the successful track record of random assignment in very diverse environments. In part building on this, there has recently been an important expansion of random assignment studies in education.[41] At a time of growing pressure to improve the performance of the nation's schools, these studies promise to bring new rigor to our understanding of the effectiveness of alternative reform strategies. It is critically important to push forward on this front.

[41] See, for example, Pauly and Thompson, 1993; Kemple, 1998; Kemple and Snipes, 2000; Cook, 1999; Nave, Miech, and Mosteller, 1998; and Mosteller, 1999.

REFERENCES

Aaron, Henry J. *Politics and the Professors: The Great Society in Perspective.* Washington, D.C.: Brookings Institution Press, 1978.

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

Betsey, Charles L., Robinson G. Hollister, and Mary R. Papageorgiou. *Youth Employment and Training Programs: The YEDPA Years.* Washington, D.C.: National Academy Press, 1985.

Bloom, Dan. *After AFDC: Welfare-to-Work Choices and Challenges for States.* New York: MDRC, 1997.

Bloom, Howard S. *Building a Convincing Test of a Public Housing Employment Program Using Non-Experimental Methods: Planning for the Jobs-Plus Demonstration.* New York: MDRC, 1999.

Boruch, Robert F. *Randomized Experiments for Planning and Evaluation.* Thousand Oaks, Calif.: Sage Publications, 1997.

Boruch, Robert, Brooke Snyder, and Dorothy DeMoya. "The Importance of Randomized Field Trials." Paper presented at meeting of the American Academy of Arts and Sciences, 1999.

Cave, George, Fred Doolittle, Hans Bos, and Cyril Toussaint. *JOBSTART: Final Report on a Program for School Dropouts.* New York: MDRC, 1993.

Connell, James, Anne Kubisch, Lisbeth Schorr, and Carol Weiss, eds. *New Approaches to Evaluating Community Initiatives: Concepts, Methods, and Contexts.* Roundtable on Comprehensive Community. Washington, D.C.: Aspen Institute, 1995.

Cook, Thomas. "Considering the Major Arguments Against Random Assignment: An Analysis of the Intellectual Culture Surrounding Evaluation in American Schools of Education." Paper presented at the Harvard Faculty Seminar on Experiments in Education, Cambridge, Mass., 1999.

Doolittle, Fred, Darlene Hasselbring, and Linda Traeger. "Lessons on Site Relations from the JTPA Team: Test Pilots for Random Assignment." Internal paper. New York: MDRC, September 16, 1990.

Doolittle, Fred, and Linda Traeger. *Implementing the National JTPA Study.* New York: MDRC, 1990.

Friedlander, Daniel, and Gary Burtless. *Five Years After: The Long-Term Effects of Welfare-to-Work Programs*. New York: Russell Sage Foundation, 1995.

Goldman, Barbara. *Impacts of the Immediate Job Search Assistance Experiment: Louisville WIN Research Laboratory Project.* New York: MDRC, 1981.

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

Greenberg, David, and Mark Shroder. *The Digest of Social Experiments.* 2d ed. Washington, D.C.: Urban Institute Press, 1997.

Greenberg, David, and Michael Wiseman. "What Did the OBRA Demonstrations Do?" In *Evaluating Employment and Training Programs*, edited by Charles Manski and E. Garfinkel. Cambridge, Mass.: Harvard University Press, 1992.

Gueron, Judith M. "The Supported Work Experiment." In *Employing the Unemployed,* edited by Eli Ginzberg. New York: Basic Books, 1980.

———. "Lessons from Managing the Supported Work Demonstration." In *The National Supported Work Demonstration,* edited by Robinson G. Hollister Jr., Peter Kemper, and Rebecca A. Maynard. Madison: University of Wisconsin Press, 1984.

———. "The Demonstration of State Work/Welfare Initiatives." In *Randomization and Field Experimentation* (special issue of *New Directions for Program Evaluation* 28 [December 1985]: 5-14), edited by Robert F. Boruch and Werner Wothke. San Francisco and London: Jossey-Bass, 1985.

———. "Learning About Welfare Reform: Lessons from State-Based Evaluations." *New Directions for Evaluation* 76 (Winter 1997):79-94.

Gueron, Judith M., and Edward Pauly. *From Welfare to Work.* New York: Russell Sage Foundation, 1991.

Hamilton, Gayle, Thomas Brock, Mary Farrell, Daniel Friedlander, and Kristen Harknett. *National Evaluation of Welfare-to-Work Strategies: 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, Administration for Children and Families and Office of the Assistant Secretary for Planning and Evaluation, and U.S.

Department of Education, Office of the Under Secretary and Office of Vocational and Adult Education, 1997.

Haskins, Ron. “Congress Writes a Law: Research and Welfare Reform.” *Journal of Policy Analysis and Management* 10(4) (1991): 616–632.

Hollister, Robinson G. Jr., Peter Kemper, and Rebecca A. Maynard, eds. *The National Supported Work Demonstration*. Madison: University of Wisconsin Press, 1984.

Kemple, James J. “Using Random Assignment Field Experiments to Measure the Effects of School-Based Education Interventions.” Paper prepared for the annual conference of the Association for Public Policy Analysis and Management, New York, October 1998.

Kemple, James J., and Jason C. Snipes. *Career Academies: Impacts on Students' Engagement and Performance in High School.* New York: MDRC, 2000.

Kornfeld, R., and Howard Bloom. “Measuring the Impacts of Social Programs on the Earnings and Employment of Low-Income Persons: Do UI Wage Records and Surveys Agree?” *Journal of Labor Economics* 17(1) (1999).

Leiman, Joan M. *The WIN Labs: A Federal/Local Partnership in Social Research.* New York: MDRC, 1982.

Miller, Cynthia, Virginia Knox, Patricia Auspos, Jo Anna Hunter-Manns, and Alan Orenstein. *Making Welfare Work and Work Pay: Implementation and 18-Month Impacts of the Minnesota Family Investment Program.* New York: MDRC, 1997.

Mosteller, Frederick. Forum: “The Case for Smaller Classes.” *Harvard Magazine (May–June* 1999).

Nave, Bill, Edward J. Miech, and Frederick Mosteller. “A Rare Design: The Role of Field
Trials in Evaluating School Practices.” Paper presented at meeting of the American Academy of Arts and Sciences, Harvard University, Cambridge, Mass., 1998.

Pauly, Edward, and Deborah E. Thompson. “Assisting Schools and Disadvantaged Children by Getting and Using Better Evidence on What Works in Chapter 1: The Opportunities and Limitations of Field Tests Using Random Assignment.” New York: MDRC, 1993.

Quint, Janet, Johannes M. Bos, and Denise F. Polit. *New Chance: Final Report on a Comprehensive Program for Young Mothers in Poverty and Their Children.* New York: MDRC, 1997.

Quint, Janet, Kathryn Edin, Maria L. Buck, Barbara Fink, Yolanda C. Padilla, Olis Simmons-Hewitt, and Mary Eustace Valmont. *Big Cities and Welfare Reform: Early Implementation and Ethnographic Findings from the Project on Devolution and Urban Change.* New York: MDRC, 1999.

Riccio, James. *Mobilizing Public Housing Communities for Work: Origins and Early Accomplishments of the Jobs-Plus Demonstration.* New York: MDRC, 1999.

Riccio, James, Daniel Friedlander, and Stephen Freedman. *GAIN: Benefits, Costs, and Three-Year Impacts of a Welfare-to-Work Program.* New York: MDRC, 1994.

Szanton, Peter L. "The Remarkable 'Quango': Knowledge, Politics, and Welfare Reform." *Journal of Policy Analysis and Management* 10(4) (1991): 590–602.

U.S. Department of Labor. *Recommendations of the Job Training Longitudinal Survey Research Advisory Panel.* Washington, D.C.: U.S. Department of Labor, 1985.

———. *What's Working (and What's Not): A Summary of Research on the Economic Impacts of Employment and Training Programs.* Washington, D.C.: U.S. Department of Labor, 1995.

Wiseman, Michael. "Research and Policy: An Afterword for the Symposium on the Family Support Act of 1988." *Journal of Policy Analysis and Management* 10 (4) (1991): 657–666.