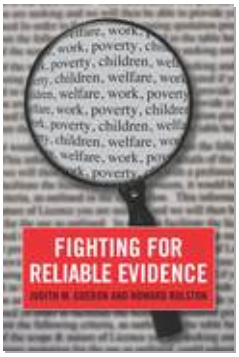
Fighting for Reliable Evidence

Gueron, Judith M., Rolston, Howard



Published by Russell Sage Foundation

CHAPTER 1

Introduction: The Issue, the Method, and the Story in Brief*

> The federal government should follow a systematic experimentation strategy in seeking to improve the effectiveness of social action programs. . . . The process of developing new methods, trying them out, modifying them, trying them again, will have to be continuous. . . . [Otherwise] it is hard to see how we will make much progress in increasing the effectiveness of our social services.
>
> —Alice Rivlin[1]

> I was struck by the power of this novel [random assignment] technique to cut through the clouds of confusing correlations that make the inference of causality so hazardous. . . . I have never quite lost my sense of wonder at this amazing fact.
>
> —Larry Orr[2]

How can we know whether social programs do what they are designed to do? Can that question even be answered convincingly? A simple example illustrates the problem:

The governor of a large midwestern state is under pressure. Welfare rolls are rising. Legislators are pushing for action. Armed with a report from a blue-ribbon panel, she recommends a new program with a catchy name: WoW, for

*Chapter 1 authored by Judith M. Gueron and Howard Rolston.

Working over Welfare. A year later the program is up and running, the rolls are down, and the press declares a winner. Basking in the glow, she runs for national office.

But did WoW actually bring about the change? Did it cause the rolls to go down, or were other factors at work? Take the case of Mary, a single mother who was laid off from her job, applied for welfare, and, under the new rules, was told she had to look for a job or enter a training program. Within six months, she was back at work. Since "people like her" living in that state typically took a few years to get off welfare, this sounds like a clear WoW success. But how can we be sure the new program was really responsible for what happened? Mary hated being on the dole. Perhaps the comparison with what "people like her" (as measured by her background and personal characteristics) had been doing did not really mimic her likely behavior under the previous system. Also, during this same period the job market in her area improved. So how can we determine whether Mary got a job because of WoW, or because of her drive to escape welfare, or because of the stronger economy? Furthermore, can that question really be answered with any confidence by looking at what thousands of Marys across the state had been doing without WoW? Or is this inevitably a case of "on one hand" and "on the other"?

This book describes how an approach that is akin to clinical trials in medicine solved this type of social policy problem—yielding a simple but utterly convincing answer to the question of whether a social program actually achieves its goals. A type of random selection similar to tossing a coin or running a lottery was developed under which similarly eligible people applying for a new form of welfare were randomly placed in either the new program or the prior system. This approach enabled analysts to measure and compare the outcomes of two groups that differed systematically only in their welfare program experience. What the new program achieved can be accurately estimated as the difference in outcomes between the two groups, because they were randomly selected from the same group and lived through the same shifting economic and social conditions. Both the claims of politicians and academic arguments about the "correct" interpretation of complex statistics were successfully replaced by the simple concept of a coin toss.

This book tells the story, step by step, of how this research strategy—which for much of the period was ignored or belittled by most scholars—was shown to be feasible in the rough-and-tumble real world and how it went on to win converts and influence people. We tell our tale deliberately as an action story,

which indeed it is, to give readers some of the intense flavor of our struggle as it unfolded. Our story focuses on the policy issue of providing financial support while encouraging self-reliance. But our overriding purpose is much deeper. Our intent is to use our experience to draw useful lessons for people struggling to isolate the impacts of initiatives in other social policy areas, as well as researchers, government officials, funders, and interested citizens who share our belief that the development of reliable evidence is central to improving social policy and practice, whatever the specific area may be.

THE ISSUE: PROVIDING SUPPORT VERSUS ENCOURAGING SELF-RELIANCE

The basic dilemma of welfare—identified at least as far back as Tudor England, when it was first tackled legislatively in the Poor Law of 1601—is how to assist needy citizens without stimulating behavior that perpetuates poverty and dependence. The U.S. cash welfare program created by Franklin D. Roosevelt's New Deal sought to avoid this conundrum by restricting benefits to a small group of single mothers who were not expected to work: primarily poor widows who, according to the general view of the time, should stay at home and care for their children rather than be forced to go to work and put the children in orphanages. Since these were cases of hardship, not choice, the issue of work incentives did not arise. (Throughout this book, we use the word *welfare* to refer to the federal-state program that provides cash assistance to low-income families. In 1935, when cash welfare replaced the existing state mothers' pensions programs, it was called Aid to Dependent Children. Subsequently, though the welfare program always provided cash assistance in some form, the name was changed to Aid to Families with Dependent Children [AFDC] and currently Temporary Assistance for Needy Families [TANF].)

Over the next sixty years, however, reality increasingly diverged from this vision, eroding support for the program: the rolls and costs grew dramatically (particularly between 1965 and 1975); the vast majority of single mothers receiving welfare were not widows but divorced, separated, or never married; and women across the country (including single parents with very young children) were flooding into the labor force, often not by choice.[3] These changes raised questions about the equity of long-term support for one group of single mothers and whether the very design of the program was having a range of unintended side effects. These potentially included encouraging fam-

ily breakup and teen pregnancy, discouraging women from earning a living, and making it easier for fathers to leave their families to avoid having to support their own children.

Once it became clear that some welfare recipients were indeed employable, the central tension epitomized in the Poor Law led to the following logical chain:

- Most people have to work for income. Welfare provides an alternative and thus reduces the incentive for people to work.
- So that only those in need receive assistance, benefits must go down as earnings go up.
- This benefit reduction rate in effect functions like a tax on individuals' earnings, further reducing the incentive for welfare recipients to take jobs.
- Since welfare benefits are financed from the taxes other members of society pay, there is always public pressure to keep program costs low by keeping benefits low and benefit reduction rates high, the latter exacerbating the negative incentives of welfare on parents' working.

The weaknesses of programs described in this way are obvious. But when they look at ways to improve them, policy makers differ, often strongly, on which objectives should be primary—cutting costs, enforcing work (even for mothers of small children?), reducing poverty, protecting children, keeping fathers in the home, or strengthening families. This disagreement is not surprising, since the choices involve often contentious trade-offs among the central value judgments of income redistribution, social justice, the roles of men and women, economic efficiency, and individual responsibility.

Important for our story, in contrast to Social Security, which was fully funded by the federal government and operated under standard, nationwide rules, the welfare program was designed as a federal-state partnership. The program was a federal entitlement, meaning that no person who satisfied the eligibility criteria could be denied benefits. The states retained substantial discretion over those criteria as well as over grant levels, however, while sharing the program cost (in varying proportions over time) with the federal government. As a result, not only the states but also the federal government would be on the hook to pay for any state decision to increase the number of beneficiaries or the level of support.

Over the ensuing years, reflecting the dramatic shift in attitudes toward single mothers and thus in the basic rationale for the program, the original New Deal welfare system was progressively replaced by a program that used various means to encourage, assist, or require an increasing share of women to seek and accept jobs as a condition for receiving cash assistance. The cycles of reform reflected battles about both the balance between competing objectives and how to achieve them and drew on emerging evidence from the experiments we describe in this book.

A major turning point came in 1956, when the federal government recognized the goal of encouraging independence by expanding AFDC to include services to help persons caring for the recipient children to "attain the maximum self-support and personal independence." In 1961 the federal government for the first time recognized the family stability goal by expanding the program to include, at state option, the unemployed parent program (AFDC-UP), under which two-parent families in which the father was employable but had become unemployed became eligible for cash support. At this point, however, AFDC program benefits were still calculated as if the program were directed solely at reducing hardship rather than also encouraging work. Thus, if recipients didn't work, they received the full cash benefit, which depended on family size. But if they started to earn, their cash benefit was reduced dollar for dollar, leaving the family no better off financially.

The War on Poverty, a part of President Lyndon B. Johnson's Great Society initiative, brought the issue of poverty to the fore. In this context, two distinguished economists from opposite ends of the political spectrum (both subsequent Nobel Prize winners) advocated a new idea: the negative income tax (NIT).[4] The negative income tax would combine the positive tax system with the welfare (cash-benefit) system, so that those above and below a given income threshold would face similar tax rate schedules. Those with no earned income would receive the maximum payment, called the guarantee. As they started earning, their guarantee would be taxed at a rate that gradually eliminated it as income rose, at which point the system would merge into the positive tax system. Central to the NIT was the idea that payments should be based on income, not a particular status (such as single parenthood), thus removing a concern about the perverse incentives on family formation and stability.

Not surprisingly, the idea of an NIT raised a firestorm of questions and concerns when it hit the arena of public debate, including concern that many

households could be made worse off under an NIT than under the then-current network of income support programs. The major question for our story was how the poor would react to the change with regard to their attitude toward work. By 1967 fewer than half the states had implemented an AFDC-UP program, and even in those states the number receiving benefits was relatively low, leaving most two-parent poor families outside the system. How much less might these parents, especially fathers, work as a result of greatly expanded eligibility for cash assistance? Or would lower tax rates lead to greater work effort?

Economic theory unambiguously predicts that extending welfare to a new group will reduce their work effort. But the effect of different tax rates is ambiguous. In particular, lower tax rates are something of a two-edged sword: they clearly encourage people who are not working to take a job; but they also keep working families, who would have become ineligible under a higher tax rate, on welfare longer and thereby extend all the negative work incentives to those families. Which effect would predominate—more work by those currently on welfare or less work by those newly affected by it? Economic theory could not say, and no reliable evidence existed to answer the question.

Early in 1967, a grant application was submitted to the Office of Economic Opportunity—the federal administrative home of the War on Poverty—proposing a field test to answer these questions. The idea was to mount a field experiment to test empirical responses to a range of NIT guarantee–tax rate combinations. The experiment would be directed to the working poor—that is, two-parent families who would be eligible for cash assistance if low income were the only criterion for eligibility, the very group skeptics of unrestricted cash assistance were most concerned would work less. The really novel idea in the grant application was its methodology. It proposed to determine the effects of an NIT by selecting a population of low-income couples and using a coin toss, lottery, or similar random mechanism to allocate them either to one of a series of experimental groups, who would receive different NIT plans (that is, differing guarantee generosities or tax rates), or to a control group, who would simply continue with their lives. Although random assignment as a way of identifying cause and effect in medical clinical trials was already in use, this was the first time that a random process had been suggested as a way to test cause and effect on a large scale in relation to potential social policy reform.[5]

The initial NIT proposal was further refined and was then implemented as the New Jersey Negative Income Tax Experiment beginning in 1968. The statistical design of the experiment was the work primarily of economists at the recently created Institute for Research on Poverty at the University of Wisconsin–Madison, a research center established by the Office of Economic Opportunity. The field operations of the experiment were designed and run by Mathematica (later Mathematica Policy Research), a research firm located in Princeton, New Jersey. No operating welfare or other public program office was involved, and no social service component was tested. The experiment began with little fanfare, until the Nixon administration pushed it into the spotlight by using and misusing some of the very early results in its campaign to push a legislative welfare reform proposal that included a type of NIT for families as its centerpiece. The Nixon proposal, called the Family Assistance Plan, passed the U.S. House of Representatives in April 1969. It then went through a long series of revisions in what had become the impossible hope of making it politically palatable to both the right and the left in the Senate. No version ever passed the Senate, which killed the legislation for the last time in October 1972.

The New Jersey experiment, which finished field operations in 1972 and published its three-volume final report in 1976–1977 (Kershaw and Fair 1976; Watts and Rees 1977a, 1977b), was followed closely over the next several years by three essentially similar experiments—the most ambitious being the Seattle/Denver Income Maintenance Experiment, known as SIME/DIME—each of which used random assignment to estimate the effects of NITs of different generosities, in different environments, extending the population to single parents, and, in some cases, offering additional services designed to help recipients get work. The bottom line from the studies was that the NIT reduced rather than increased overall work effort.[6]

This whole body of research and practice—referred to collectively as the income maintenance experiments—might well have remained a relatively minor footnote in the history of U.S. social policy were it not for its legacy: first, a group of analysts (in academia, research organizations, and government agencies) who learned their craft in one or more of the income maintenance experiments, participated in important ways in many of the experiments we describe in subsequent chapters, and were fearless advocates for our cause when we needed them; second, the body of knowledge and experience about

how to design and implement a random assignment experiment that these analysts brought with them and on which we were able to build; third, an important negative lesson about when and how to release findings into a politically charged atmosphere. This legacy is an invaluable part of our story.

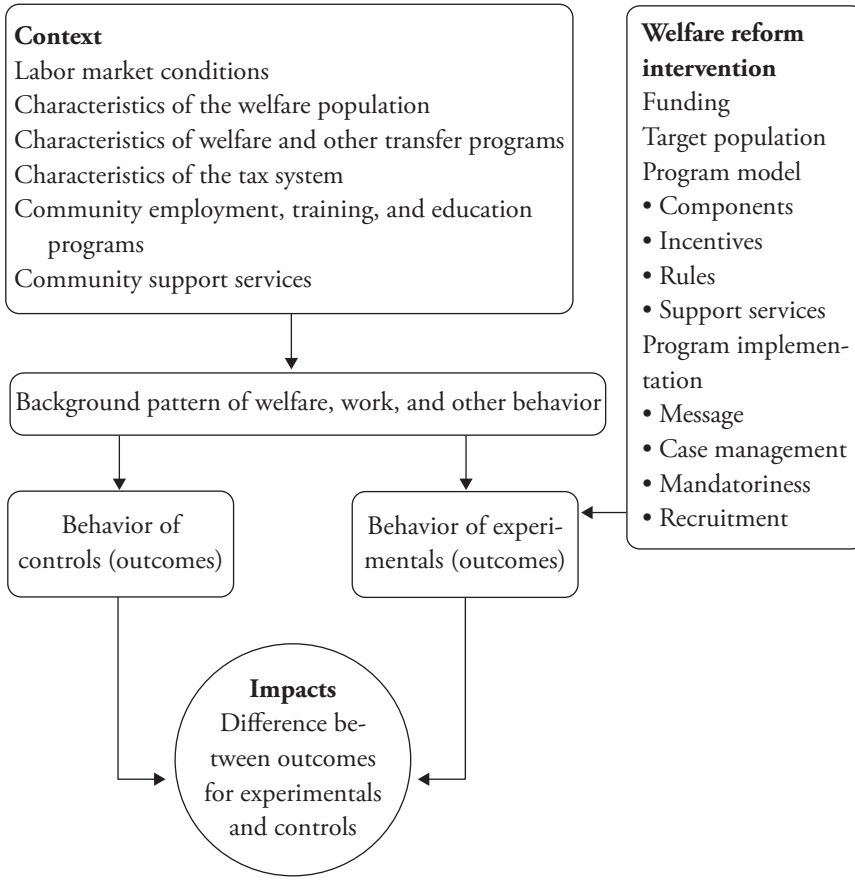
Although Congress failed to pass an NIT, it did during these years enact changes that to this day remain key tools used by states to affect behavior: reducing tax rates for AFDC and establishing a welfare-to-work program (originally called the Work Incentive, or WIN, program) under which employable welfare recipients (a group first thought to be small but redefined upward repeatedly over the next thirty years) could be required to work, or participate in activities intended to help them get jobs, or face financial penalties. The hope was that either the assistance or the threat of the loss of benefits for noncooperation would spur people to take a job.

THE METHOD: RANDOM ASSIGNMENT IN THE CONTEXT OF A SOCIAL EXPERIMENT

The fundamental concept is familiar from medical research. We have all seen headlines like "Randomized Clinical Trial Challenges Conventional Wisdom." Whether the issue is hormone replacement therapy, colon cancer drugs, or a new heart valve, the logic is the same. Evidence from one or many clinical trials overturns long-standing practices based on observational studies or clinical experience. The randomized trial in medicine, in which a group of people with defined characteristics that make them eligible to participate in the trial are allocated at random either to one or more experimental groups, which get the test treatments, or to a control group, which gets the prevailing level of care or a placebo. The impact of the treatment is then calculated as the resulting difference in outcomes, along the relevant dimensions, between the two (or more) groups. Because, owing to random assignment, the nontreatment characteristics of the groups can be assumed not to differ in any systematic way that would affect success, any difference in outcomes can be ascribed with a known degree of statistical confidence to the treatment. All other influences on the outcomes are washed out in this calculation because they are the same for the control group and however many experimental groups there may be.

In a field or social experiment, just as in a medical trial, people are randomly placed into two (or several) groups: one (or more) involved in the program or programs being tested (the treatment or experimental group[s]);

Figure 1.1 Factors Affecting the Impacts of Welfare Reform Programs



Source: Authors' adaptation based on Gueron and Pauly (1991, figure 2.1).

and one receiving no special attention (the control group). If well implemented (an extremely big "if" that we return to repeatedly in later chapters), any subsequent difference in behavior among the people in the different groups (behavior that can affect a program's benefits and costs) can be attributed to the treatment to which each group is exposed. Figure 1.1 sketches out the types of environmental or personal factors controlled for in this conceptually simple way.

An alternative way of explaining the concept of random assignment is to

talk directly about the counterfactual—what would have happened without the program. There is no way to observe this directly because the program does, in fact, exist. The beauty of random assignment into otherwise identical groups is that it allows measurement of the counterfactual, which is the experience of the control group as it moves through time. If the economic environment suddenly improves, for example, it improves for the control group in exactly the same way that it improves for the treated group. The transparency of the method—no complex statistics, scant potential for researcher bias—is obvious.

This sounds wonderfully simple, and in concept it is. But as our story in subsequent chapters makes clear, how it is put into practice—how the control group is in fact insulated from the program experience; and when, in the process of enrolling for the program and beginning to be affected by it, the random assignment allocation actually takes place—is extremely complex. These issues define the exact question the evaluation will answer and where, as the experiment gets further and further integrated into ongoing program operations, the random assignment process can best be carried out without destroying either the rigor of the experiment or the operations of the program being evaluated.

Over the period covered in this book, the dimensions of experiments in this area of social policy changed in numerous ways: the background conditions, level of control, scale, and subject. They moved from tests of voluntary opportunities to those of mandatory obligations; from pilots for several hundred (and then thousands of) people to evaluations of full-scale programs involving tens of thousands; from centralized direction, funding, and leverage to a more decentralized partnership with states and localities; from tests of stand-alone programs to tests of components within the welfare system to tests of multidimensional systemwide reforms. Substantively, they expanded from welfare-to-work programs only to tests of financial work incentives, time limits on welfare, comparisons of education-first versus jobs-first approaches, child support enforcement, and a wide range of other program variants designed to change the behavior of recipients in particular ways. And they expanded beyond welfare itself to include the broader population at risk of welfare recipiency but not currently receiving assistance, as well as to nontraditional strategies such as promoting healthy marriage among low-income unmarried and married couples. The substantive issues the experiments sought

to clarify were all causal questions, which changed as the results from previous experiments and other research added to the cumulative knowledge base.

As noted, many of the test programs were complex and multidimensional, and it quickly became clear that defining the "treatment" involved understanding how and whether the program was actually implemented. Reform legislation might sound simple—for example, change welfare from a system that primarily paid checks to one that also imposed work and participation obligations and penalized people who balked. But state and local staff might not be willing or able to actually implement the policy change, especially given resource constraints and competing pressures to protect particular groups. Understanding the nature, feasibility, and replicability of the reform programs—the what, how, and why questions—thus became integral to interpreting experimental results.

THE STORY IN BRIEF

When our story began, we did not imagine we were launching a forty-year adventure to test the feasibility and potential of the random assignment approach to learning. There was no overarching master plan; nor did any of us anticipate that the hurdle would keep getting higher, either because political changes would create an increasingly demanding environment or because we would seek to address more—and more complex—issues. We did not envision or prepare for the battles ahead. What we did have were questions. We started, in our relative ignorance, with no answers to even the most basic questions about different strategies to move people from welfare to work: Do they have any effect? For whom? At what cost?

Despite ignorance about what would work, the pressure was on to change welfare. As successive presidents and governors promoted highly visible and often highly controversial ways to fix what they saw as a failed system, different actors—at different times and for different reasons—sought more reliable evidence on whether the proposals would deliver on the claims. The problem was not a lack of research. There had been plenty of demonstrations, evaluations, and studies of welfare and employment and training programs, but often they had ended in unresolved arguments about methodology and credibility. To get stronger proof of effectiveness, an initially small group of people pushed to try out random assignment.

Our book focuses on a subset of the scores of experiments in which one or

both of us played a direct role. Although they were by no means the only ones that followed the original, pathbreaking income-maintenance experiments, they illustrate how social experiments in the welfare area moved out of the researcher-controlled environment (testing behavioral responses with statistical designs driven largely by economic theory) into the more complex context of mainstream public agencies and became the federal standard for the evaluation of real-world programs. In the process, we tell why and how MDRC and different agencies within the U.S. Department of Health and Human Services (HHS) became committed to this approach and acted to fashion a coherent knowledge-building agenda, how people inside and outside government sparked and sustained this innovation, how the experiments were successful in answering important policy questions, and how the findings affected Federal and state policy. The rest of chapter 1 summarizes the highpoints of our story and the lessons learned.

Act One: Chapters 2 and 3

During the first act of our story, random assignment was almost an afterthought grafted onto new initiatives launched by private and public sector social entrepreneurs. The initial step was a specially funded demonstration run by community-based organizations (chapter 2). Supported Work—a project initiated by the Ford Foundation and managed by MDRC (specially created for the purpose), with funding also contributed by six federal agencies led by the Department of Labor—fit into the try-small-before-you-spend-big vision of policy making. The idea was to run a relatively small demonstration (several hundred volunteers per site) of a specific program model (structured paid jobs to help transition hard-to-employ people into the regular labor market); to randomly assign a total of 6,500 people in four very disadvantaged target groups (former prisoners, former addicts, AFDC mothers, and unemployed youth) to get the strongest possible evidence on whether the program affected key outcomes and how benefits compared with costs; and only then to recommend whether it was worthy of broader replication.

When the project started in 1974, many thought it would implode—that program operators would refuse to let an outsider running a lottery decide who could and who could not get in. Surprisingly, that did not happen. Through a combination of persuasion, flexibility on the noncore issues, and absolute rigidity on the central issue of random assignment, the team at MDRC, Mathematica Policy Research, and the Institute for Research on

Poverty—a team that included many veterans from the original New Jersey NIT experiment—was able to convince organizations to sign up. At the time, we concluded that success was the result of four circumstances: more people than could be served volunteered for the new program and study, making it relatively easy to argue that a lottery was a fair way to ration scarce opportunities; the project was tightly controlled from the center; local sites were offered millions of dollars if they agreed to play by the rules of evidence-based science; and the intentionally low profile slipped the experiment under the political and press radar.

The next step came in 1977, when the national director of the WIN program proposed creating "laboratories" in a few offices to test locally generated ideas to improve performance (chapter 3). With his sponsorship, it proved feasible to insinuate random assignment into the intake process in regular government offices without unduly disrupting operations. Although recruiting sites and putting the procedures in place took patience, skill, and obstinacy, there were no knock-down, drag-out fights. Two of the Supported Work conditions continued—the tests were voluntary, and the programs were small and largely invisible to the press and public—but there were no big bucks and no tight central control. Instead, there was a new twist: creation of programmatic and research partnerships among federal agency staff, local WIN practitioners, and evaluators, with the hope that the arrangement would lead to smarter innovations and stronger research (that is, both better questions and more support for the research requirements and findings).

Act Two: Chapters 4, 5, and 6

In the second act, experiments moved into real-world, public-private partnerships. Ronald Reagan's election in 1980 transformed welfare policy and research. All the favorable conditions from Supported Work vanished. The new administration proposed legislation that imposed new obligations on welfare recipients and significantly restricted benefit eligibility—at the same time that it dramatically shrank the federal funds available for testing programs or for policy research more generally.

But there turned out to be an unexpected saving grace in what emerged from Congress with respect to work programs: newfound flexibility for state welfare agencies to undertake state-based initiatives to move welfare recipients into work. Reeling from the cancelation of multiple studies (and after letting go 45 percent of its staff), in desperation MDRC invented a partnership-with-

states paradigm that, ironically, ended up having greater relevance and policy impact than earlier experiments and became the model that flourished for the next thirty years.

This strategy (chapters 4 and 6) drew on both the WIN Laboratory vision and the 1981 federal budget bill that gave state agencies the option to change and take ownership of welfare-to-work programs. (States could require single mothers to work in return for their benefits in so-called workfare or community work experience programs and could restructure WIN, which at the time was judged ineffective in imposing a participation requirement.) MDRC's key insight was to make a reality out of the cliché that the states were laboratories. The idea was to graft experiments onto the often controversial demonstrations that emerged as governors responded enthusiastically to the opportunity to take control of WIN and, to a lesser extent, to operate community work experience programs. Instead of testing a uniform model in multiple sites (as in Supported Work), the resulting Work/Welfare demonstration used eight state-specific experiments to assess programs that reflected each state's particular values, resources, goals, and capabilities—but primarily required people to search for a job—with random assignment integrated into the helter-skelter of normal public agency operations. The initial reaction of welfare commissioners to this idea was incredulity: Is this ethical? Will it impede operations? Will it explode?

MDRC ultimately convinced administrators by making three promises. The first was to ensure that the evaluation would answer states' questions, inform their staff, improve their programs, meet high standards, avoid political minefields, not hinder performance, provide a spot in the limelight, and satisfy the then vague evaluation requirements imposed by the federal government in return for permitting states to make innovations beyond what was allowed by law. The second promise was to create a learning community for senior staff. The third was to provide real-time advice on program design.

The two additional ingredients needed to make this vision a reality were money to pay for the research and the assurance of an initially hostile Reagan administration that it would not obstruct the evaluation. In a show of creative and enabling philanthropy, the Ford Foundation put up a large challenge grant to cover half the projected cost; and federal officials (often working behind the scenes) offered significant support, including facilitating the use of matching AFDC administrative and special demonstration funds to help pay for the evaluation, with the rest coming from a few states and other foun-

dations. The result surprised even those who proposed this departure: 28,500 people were randomly assigned in eight states, no state dropped out, and there were no palace coups or citizen revolts. The success of the Work/Welfare demonstration raised the profile of and made important allies for random assignment, not only because the project demonstrated its feasibility in a more real-world context than previous experiments but also because (aided by its timing, important findings, and aggressive communications strategy) it showed that diverse audiences could appreciate the exceptional quality of the evidence and would conclude that it had an almost unprecedented influence on policy and practice.

While MDRC was implementing the Work/Welfare demonstration, a new force for experimental evaluation began to emerge in the Department of Health and Human Services (chapter 5). The Office of the Assistant Secretary for Planning and Evaluation (ASPE) in HHS had followed the Office of Economic Opportunity in sponsoring additional income maintenance experiments, had also launched the major health insurance experiment, and had helped fund and provide waivers for the Supported Work program. But it was another HHS office, the Office of Family Assistance (OFA) in the Social Security Administration, that took on this role. With no direct experience in funding or overseeing experimental evaluations, OFA launched a series of projects, one of which resulted in the addition of two sites in the Work/Welfare demonstration. In addition, OFA quietly and indirectly funded several other states that participated in the demonstration and undertook several other experiments as well. Through these experiences, OFA staff began to understand better what was necessary to design, fund, and manage experimental research and became ever more convinced of both its unique value and its feasibility.

Act Three: Chapter 7

In 1987 a new federal thrust for experimental evaluation emerged from an unexpected quarter. The federal government had the authority to grant states waivers (through section 1115 of the Social Security Act) of most AFDC program requirements in order to try program innovations in their state. Hitherto, the Reagan administration had approved section 1115 waiver demonstrations only for projects that implemented changes in line with its policy preferences. In 1987, however, the Office of Policy Development in the White House pushed HHS to interpret this authority more flexibly, encour-

aging states to apply for AFDC and other waivers to implement demonstrations of their own choosing. Alarmed that this policy would drive up the federal costs of AFDC and other means-tested open-ended entitlements, the federal Office of Management and Budget conceived of and fought internally within the administration for an approach that would require states, in return for federal flexibility, to incorporate random assignment in the design of any demonstrations for which states needed waivers. The stated purpose of this requirement (termed the quid pro quo) was to include costs in their list of impacts to be measured, so that the federal government could limit its share of program expenditures to an amount extrapolated from the costs incurred by the control group. Internal battles over the decision continued for several years, during which the federal government did not uniformly require the quid pro quo for waiver requests. But HHS and the Office of Management and Budget staff continued to fight for it. As a result, in 1992—at the beginning of an explosion of state waiver requests—random assignment became the de facto as well as the de jure standard for federal welfare reform evaluations.

Act Four: Chapters 8, 9, and 10

The fourth act involved tests of full-scale programs and, in some cases, state-initiated experiments. In that spirit, between 1986 and 1990 senior officials in California and Florida, seeking reliable information but not driven by the need for waivers, invited MDRC to use random assignment for the first time to determine the effectiveness of full-scale and ongoing programs (chapter 8). The scale of these initiatives was huge in comparison with previous random assignment efforts (the California Greater Avenues for Independence study alone included 35,000 people), attracting high visibility and great antagonism in certain quarters. Some state legislators in both states tried to close down the studies and even to impose an outright ban on control group research, threatening both projects in their entirety. Both were eventually saved, but only as a result of a coalition of supporters, led by champions in the California and Florida welfare agencies and legislatures and including advocates, congressional staff, and a small number of academics.

The year 1988 saw passage of the Family Support Act, federal welfare reform legislation that included passage of the Job Opportunities and Basic Skills Training (JOBS) program (which replaced WIN with a program that extended the participation obligation to mothers with younger children; set

minimum participation standards; and, like the Greater Avenues for Independence program, emphasized education). A major hypothesis underlying JOBS was that remediation of basic education deficits was central to improving employment outcomes for potential long-term AFDC recipients. Relying on the fact that the legislation could be interpreted as requiring HHS to conduct a random assignment study to evaluate at least some components of the legislation, HHS staff, jointly in the Family Support Administration (a new agency that included OFA) and ASPE, took the opportunity to initiate a competition for an impact evaluation of the new program. This became the JOBS evaluation, the contract for which MDRC won.

As with MDRC's post-1986 evaluations, the JOBS evaluation assessed large-scale, operating programs that state and local governments volunteered to put under the microscope (chapter 9). Together, HHS and MDRC designed an evaluation that was able to explore in depth the underlying basic education hypothesis of the JOBS program and, by creating stylized labor-force attachment programs that were experimentally compared with stylized human-capital development programs in head-to-head tests, greatly strengthened the findings. Since Greater Avenues for Independence became California's JOBS program, the two studies together provided even stronger evidence than either could have done alone. As MDRC was the evaluator of both, common measurement across the two studies enabled powerful syntheses of the results to emerge in later years. Finally, by the end of the 1990s, after nearly two decades of capacity building, the JOBS evaluation and its other work would again propel HHS to the forefront of federal leadership of experimental social policy research.

By the early 1990s, the credibility and prestige of the earlier random assignment studies (and by extension of the states that participated in them), and the growing evidence of the feasibility of this approach and the weakness of alternative methods, prompted the Canadian government, the state of Minnesota, and the New Hope program in Milwaukee to seek experimental evaluations of programs intended to make work pay more than welfare (chapter 10). Because these and the simultaneous JOBS evaluation assessed impacts on poverty, employment, dependency, costs, children, and other family outcomes, MDRC and its partners were able to assess the trade-offs in results (including the impacts on children) between these different strategies, producing compelling syntheses that combined results from multiple studies. Although implementing each study brought challenges, there was growing ac-

ceptance that if one wanted evidence that would both withstand the scrutiny of the research community and have an impact on policy, this was the required route.

Act Five: Chapter 11

The fifth and final act in our story began with passage of the Personal Responsibility and Work Opportunity Reconciliation Act (PRWORA) in 1996, the federal reform law that replaced AFDC with TANF. The major provisions of TANF included giving states enormous flexibility to design their programs; potentially penalizing states that did not require significant percentages of adults on welfare to participate in work activities; imposing a lifetime limit of five years on federal welfare receipt; and replacing open-ended federal matching of state dollars with a block-grant formula that did not depend on caseload size or state costs. From now on, states could expand their cash assistance programs, but the federal government would not have to share the cost of any expansion. In President Bill Clinton's famous words, TANF would "end welfare as we know it."

Important for our story, PRWORA eliminated the waiver quid pro quo, under which the government had required that states perform a random assignment evaluation in return for flexibility to test new welfare reform initiatives. Many observers were understandably concerned that this would be the end of the random assignment era. For several reasons, this turned out not to be the case. First, Senator Daniel Patrick Moynihan and key congressional staff inserted language into TANF that provided money for HHS to continue to fund studies, as well as strong authorizing language. This was critical, because the new block-grant funding scheme meant that any program evaluation costs would otherwise have to come directly out of state funds (from either their share of TANF's block-grant funds or state general revenues). Federal funding was particularly important, as many foundations either shifted to funding observational studies or stopped welfare evaluation funding altogether. Second, the Administration for Children and Families (ACF), the successor to the Family Support Administration, built on earlier lessons to create new partnerships.

ACF was able to combine the new funds with others to help finance a significant number of demonstration studies with willing states and, in a few years, to initiate multisite experiments in new employment areas of interest to states. In some instances ASPE was a partner in these new ventures. In the first

years after PRWORA, the new research funds were critical in enabling ACF to support to completion a subset of the best experimental state-waiver studies. Building on these continuing waiver experiments allowed ACF to foster additional contributions to the syntheses of results.

After this initial effort not to lose the investment previously made in the waiver experiments, ACF, working with states, launched two major experimental projects aimed at addressing new questions that the policy environment of TANF—stricter work requirements and time limits—generated. The first, the Employment Retention and Advancement demonstration, went beyond programs designed to move welfare recipients into work and aimed at identifying effective strategies for helping them stay in jobs and advance in the workforce. The second project, the Hard-to-Employ demonstration, focused on finding effective approaches for helping individuals with significant work-limiting problems to be more successful in the labor market. Thus both studies examined programs that moved beyond the job search, work experience, and remediation welfare-to-work programs. ACF and the Department of Labor have continued to launch major experimental evaluations of employment interventions for low-income populations.

OVERARCHING LESSONS FOR BUILDING RELIABLE EVIDENCE

The forty-five-year experimentation with random assignment showed that this technique could be used to address most of the policy options in a wide range of conditions and, furthermore, that the distinctive quality of the evidence was often recognized and valued. This experience—which helped transform an idea that was outside the policy research mainstream into the method of choice for assessing welfare programs—provides lessons for other fields seeking similar rigor.

Our concluding chapter (chapter 12) details the many lessons from our story, including those that apply primarily to the welfare field. Here we briefly highlight the lessons that stand out for people interested in applying our experience to their particular policy area.

Methods

Embedding random assignment in real-world programs and maximizing the yield of such studies requires a blend of policy insight, technical research skills, and operational expertise. It is crucial to balance research ambition and

program needs. In many areas, the two may conflict, as they did in the welfare case. Moreover, the possibility that the pattern of impacts may not match expectations based on theory or intuition needs to be kept firmly in mind. In such cases, only a transparently credible evaluation strategy will yield evidence that convinces. Finally, building knowledge step by step, by combining experimental with nonexperimental methods and synthesizing results across studies, clarifies the how and why questions left by previous studies in ways that are much more convincing in total than would be the sum of the various parts.

Program Effectiveness

Outcomes (for example, how many people took jobs) proved again and again to be poor proxies for impacts (for example, how many took jobs as a result of the program), raising challenges for program managers who seek easily obtainable standards to drive and reward positive performance. In addition, although many reforms beat the status quo, the gains were generally modest. Economic and other conditions, and personal characteristics, were the primary influences on behavior, pointing to the importance of setting realistic expectations. Finally, different policies were more or less successful in reaching different goals—saving money, reducing poverty, or reducing dependency—suggesting the likelihood of trade-offs.

A Comprehensive Body of Evidence

The power of the welfare experiments flowed from their logic, consistency of findings across multiple locations, and relevance, reflecting the independent determination of both MDRC and HHS that successive experiments be accretive (substantively and methodologically) and respond to the dynamic policy context. The paradigm of partnership with states, though forged out of necessity, had the important benefit of producing results relevant to the diverse context of the real world, rather than seeking to identify a single, most effective model. Beyond this, the central factor that sustained the uninterrupted decades of experiments we describe was the conscious creation of a community of researchers, public officials, funders, advocates, and state and federal legislative staff who recognized and valued the distinctive quality of the evidence they produced.

Advocates for experiments need to keep in mind that research evidence—no matter how reliable—is at best a relatively small element in what are usu-

ally very political policy debates. The random-assignment welfare experiments (combined, importantly, with forceful but evenhanded marketing of the results) produced uncontested findings, widespread media coverage, and a perception that the results had an unusually strong impact on policy and practice. This does not mean the findings determined policy. They did not, nor should they. But does the experience show that uncontestable findings can be weighed in the political balance? Yes.

High-quality research (experimental or not) costs money, and the welfare area benefited from four unusual conditions: long-term support by the Ford Foundation to pursue studies of little initial interest to the federal government; creation of a federal incentive (the waiver process) for states to evaluate their programs; the entitlement financing structure that allowed states to draw down matching funds for administrative functions (which provided an open-ended subsidy for state program evaluations); and continuing congressional funding for federally sponsored evaluations. Since 1996, the first three ended, making funding more centralized and vulnerable to changes in the political winds.