The Politics and Practice of Social Experiments: Seeds of a Revolution

Judith M. Gueron
President Emerita, MDRC

March 21, 2016

The author thanks Lawrence Katz, Esther Duflo, Howard Rolston, and John Hutchins for helpful comments on an earlier draft of this chapter. The views expressed herein are the author's and do not necessarily reflect those of MDRC.

Forthcoming in the *Handbook of Field Experiments,* edited by Abhijit Banerjee and Esther Duflo.

The Politics and Practice of Social Experiments: Seeds of a Revolution

Judith M. Gueron, March 21, 2016

Between 1970 and the early 2000s, there was a revolution in support for the use of randomized experiments to evaluate social programs. Focusing on the welfare reform studies that helped to speed that transformation in the United States, this chapter describes the major challenges to randomized controlled trials (RCTs), how they emerged and were overcome, and how initial conclusions about conditions necessary to success — strong financial incentives, tight operational control, and small scale — proved to be wrong. The final section discusses lessons from this experience for other fields.

Why Focus on Welfare?

Substantive and personal reasons explain my focus on welfare. It is the field of social policy research that pioneered large-scale RCTs and in which they have had the longest uninterrupted run (almost 50 years). Many view these evaluations as having had an unusual impact on legislation, practice, research methods, and the current enthusiasm for evidence-based policy (Angrist and Pischke 2010, 5; Baron 2013, 2; de Parle 2004, 111; Greenberg, Linksz, and Mandell 2003, 238; Haskins 2006, 11; Manzi 2012, 181). The second reason is more parochial: I know this history firsthand and can provide an insider's perspective on why and how the art that sustained RCTs developed.

Although numerous books and articles present findings from or describe how to design experiments,[1] my task is different: to lay out what it took to move them from the laboratory into the real world of social programs. In doing so, I draw, often directly, from *Fighting for Reliable Evidence* (Gueron and Rolston 2013), which centers on MDRC (formerly, the Manpower Demonstration Research Corporation) and the United States Department of Health and Human Services (HHS), the two organizations that played outsized roles in shaping this story.[2] The focus on HHS (a direct or indirect funder of most of these studies) is obvious; that on a private, nonprofit company makes sense because over a critical twenty years that organization conducted many of the major evaluations and, with HHS, shaped the research agenda. Although in what follows I have sought to be objective and draw on a vast archive of contemporaneous documents and subsequent interviews and publications, I am not an impartial observer. I was an actor in these events, first as MDRC's Research Director (1974-1985) and then as its President (1986-2004).

[1] For example, see Bloom 2005, Bloom 2008, Gerber and Green 2012, Glennerster and Takavarasha 2013, Greenberg and Shroder 2004, Grogger and Karoly 2005, Gueron and Pauly 1991, Gueron and Rolston 2013, and Orr 1999.

[2] This chapter uses "HHS" as shorthand for shifting subdivisions within the agency, including the Office of Family Assistance in the Social Security Administration, variously titled offices in the Family Support Administration and the Administration for Children and Families, and the Office of the Assistant Secretary for Planning and Evaluation.

This chapter does not cover the scores of relevant studies, but highlights the turning points in a tale in which successive experiments built on the lessons and success of prior ones. Gueron and Rolston (2013) provides the details behind the headlines, including the crucial role played by particular entrepreneurs and supporters and the limited importance in the most influential evaluations of the federal policy of requiring random assignment as a condition for granting states flexibility to reform welfare.[3]

Why Experiment?

To varying degrees, the proponents of welfare experiments at MDRC and HHS shared three mutually reinforcing goals. The first was to obtain reliable and — given the long and heated controversy about welfare reform — defensible evidence of what worked and, just as importantly, what did not. Over a pivotal ten years from 1975 to 1985, these individuals became convinced that high-quality RCTs were uniquely able to produce such evidence and that there was simply no adequate alternative. Thus, their first challenge was to demonstrate *feasibility*: that it was ethical, legal, and possible to implement this untried — and at first blush to some people immoral — approach in diverse conditions. The other two goals sprang from their reasons for seeking rigorous evidence. They were not motivated by an abstract interest in methodology or theory; they wanted to inform policy and make government more effective and efficient. As a result, they sought to make the body of studies *useful,* by assuring that it addressed the most significant questions about policy and practice, and to structure the research and communicate the findings in ways that would increase the potential that they might actually be *used*.

These three goals took shape over time, in part opportunistically and in part strategically, as the conditions that had nurtured the earliest experiments disappeared. The result was an agenda of increasingly audacious RCTs — a ratcheting up in scale (from pilots for several hundred people to evaluations of full-scale, statewide reforms involving tens of thousands), in complexity (from tests of stand-alone programs to tests of multidimensional system wide reforms using multi-arm experimental designs), and in the hostility of the context (from testing funded and voluntary services offered by special programs to mandatory obligations in mainstream public agencies). Each of these steps, in turn, raised new controversies and objections and reduced centralized control. This agenda, and a resistance to conducting one-off studies or to evaluating interesting but not central issues, helped demonstrate the feasibility of RCTs under increasingly demanding conditions.

This chapter recounts how the challenges, practices, and lessons evolved in response to the shifting political, funding, and programmatic context, the knowledge gains and goals, the acquired experience, the evidence of feasibility, and the reactions to the findings. It also shows

[3] The view that clout from the federal waiver authority (what came to be called the welfare waiver quid pro quo) explains the flourishing of RCTs is a mistaken one. (For a more detailed discussion, see footnote 18.)

how the three goals became mutually reinforcing: the more the findings proved useful and used, the greater the likelihood that the relevant actors would agree to the demands of quality.

The Story

In the 1970s, knowledge about efforts to move people from welfare to work could be accurately described as in the dark ages, with no answers to the most basic questions about whether reforms had any effect, for whom, and at what cost. The prevailing mood was skepticism. The problem was not a lack of evaluations, but that studies of effectiveness all too often ended with experts gathered around a table debating methodology, an outcome that not only was the kiss of death for having an impact on policymakers but also fed the conviction that this research was just another form of advocacy and not "scientific."

The main obstacle to obtaining persuasive evidence of effectiveness comes from the reality that people on welfare do not stand still waiting for some program to give them a helping hand. Many factors influence behavior. When a woman gets a job, for example, how can one tell if it is because of the help she received, the economy improved, she got her children into day care, she simply hated the stigma and hassle of public assistance, or some combination of these or other reasons? Is it possible for an evaluation to answer this question convincingly? Can it sort out the effect of one intervention from the web of other factors? Because of this reality, the "outcomes" for people enrolled in an activity (for example, the number of individuals who get a job, earn a diploma, or leave welfare) may accurately tell you their status but will not tell you the change in status that the program caused, what researchers call its value added or "impact." The logic is clear: if some people move from welfare to work on their own, outcomes will overstate impacts. But, by how much?

To answer that question, one needs a "counterfactual," a reliable measure of what the same people would have done without the intervention. During the 1970s, researchers tried various strategies to mimic this "what if" behavior. They compared the conduct of participants with their own actions before they enrolled, or with that of people who resembled them on measured characteristics but did not volunteer, were not selected or served, or lived in a different but similar community. The main weakness of such designs was "selection bias," the risk that people in the comparison group would differ in some systematic but unmeasured and influential way from people in the experimental treatment. If selection bias occurred, the context, motivation of people in the two groups, or both would not be the same, and a comparison of their subsequent outcomes would produce a biased estimate of the program's impact.

The unique strength of random assignment is that it both solves the problem of selection bias and is transparent. Since eligible people are assigned by chance to the treatment or control group, there is no systematic difference in the groups or in the conditions they face initially or over time. If the numbers are large enough and the study is done well (two big "ifs"), the result is the right answer. On transparency, RCTs allow researchers to estimate impacts using arithmetic. Basically, all one has to do is calculate the average behavior of people in the two groups after random assignment and subtract. There may be some straightforward adjustments (which rarely affect the basic findings), but no fancy statistics, no mumbo jumbo of arcane expertise, and scant potential for researcher bias. Everyone could — and did — understand this simple process.

But the question whether it was feasible remained. In the 1960s and 1970s, researchers knew about random assignment, but most saw it as a laboratory tool that was not a realistic means to address important problems in everyday conditions. By the early 2000s, it had become clear that it was, and uniquely credible. It was also increasingly clear that alternatives would not reliably produce the right answer or make evident when they did and did not. How this change happened was not the result of some decades-long master plan, but of the iterative actions of entrepreneurs inside and outside of government.

The chapter tells the story of these individuals' push to determine causality. It does not focus on a simultaneous and coordinated effort that was of equal importance: the attempt to find out how and why programs succeeded or failed. This effort included documenting the extent to which the test treatments were implemented (their operational achievements) and determining (using varied methods) why they did or did not achieve their goals and what changes would make them more effective (Gueron and Rolston 2013, 58-59, 291, 426).

Major Challenges

Implementing a high-quality RCT means overcoming numerous obstacles:[4]

1. Gaining the initial and ongoing cooperation of the relevant administrators and organizations (including their frontline staff) with conducting intake via a lottery, defining and sustaining a distinct treatment, enforcing the research groups (which usually means not helping the control group members) initially and over time, enrolling an appropriate and adequate sample, and cooperating with various research protocols
2. Securing funds for the research and, sometimes, the test program, especially if it is a special demonstration
3. Obtaining the cooperation of the research subjects
4. Acquiring reliable and comparable data for people in the program and control groups in order to track outcomes for a long enough time to detect key effects
5. Meeting high ethical and legal standards
6. Assuring that the operating program has a fair test, in particular, that it has moved beyond the start-up phase
7. Getting all the details right and keeping the endeavor on track for the years necessary to determine potential effects.

[4] During the years discussed in this chapter, almost all of the studies involved the random assignment of individuals, not intact groups or clusters.

The first challenge is the most fundamental. The researcher needs the cooperation of people in the agencies involved. But what is in it for them? Success hinges on an ability to assure them that this approach — which for some evokes horrific images of "experimenting" with human beings — is ethical, legal, and actually necessary (that is, that a less intrusive and possibly less expensive design would not do just as well). In the 1970s and 1980s, this was a tough sell. There was limited academic support and plenty of vocal naysayers, including high-powered econometricians (who claimed that they could solve selection bias via statistical modeling or alternative designs) and researchers from diverse disciplines who argued that experiments addressed limited or secondary questions (Gueron and Rolston 2013, 270-272, 455-468). This skepticism was before newspapers routinely reported how randomized clinical trials in medicine overturned long-standing practices based on observational studies and before it had become almost trite to say that correlation did not imply causation.

As a result, the risk-to-reward calculation was stacked against experiments. Why would any politician or administrator chance adverse publicity, a potential lawsuit, bureaucratic backlash, or even staff revolt? The trick was to somehow persuade people that the benefit from being involved in the RCT exceeded these obvious dangers and that, as a result, they wanted you as much as you wanted them. To gain this cooperation, managers of randomized experiments needed to create a win-win situation. As shown in the rest of this chapter, they employed diverse tools that drew on operational, research, and political skills and savvy — a combination that I have elsewhere called an art (Gueron 2002, 32). By these means, MDRC and others were able to reverse incredulity and get many to agree to join and, in some cases in later years, even seek out participation in such studies.

Demonstrating Feasibility: the National Supported Work Demonstration

Starting in 1975, the first large random assignment study of a multisite employment program, the National Supported Work Demonstration, offered a year of carefully structured, paid work to hard-to-employ people — former prisoners, former addicts, young school dropouts, and single mothers who were long-term recipients of welfare (at the time called Aid to Families with Dependent Children [AFDC] and now Temporary Assistance for Needy Families [TANF]).[5] The hope was that participants would develop some combination of habits, attitudes, self-worth, skills, and credentials that would produce a long-term increase in employment and reduction in criminal activities, drug abuse, or welfare receipt.

Even though the country had already successfully launched several path-breaking social experiments — the negative income tax (NIT), health insurance, and housing allowance demand experiments in the 1960s and 1970s — those experiments tested variations in economic incentives: treatments that could be defined by a small number of parameters (guarantee levels, tax rates, coinsurance requirements, and so on) and that were tightly controlled and

[5] AFDC, the federal-state cash welfare program created by Franklin D. Roosevelt's New Deal, was replaced by TANF in 1996. Although the Supported Work demonstration included welfare recipients, it was viewed as a highly targeted employment program, not as a pre-test for welfare reform (Gueron and Rolston 2013, 29).

administered by the researchers. The Supported Work challenge promised to be harder, with much less researcher control, and included convincing 10 mission-driven, community-based nonprofit organizations to operate a complex program and use an intake lottery. With a 45-year track record of success, it is easy to get blasé, but at the time random assignment in such a context was unheard of. The message was clear: it simply cannot be done. Program operators would implacably oppose turning people away based on some random process. The approach would be viewed as cold hearted, immoral, and akin to asking a doctor to deny a patient a known cure.

Given the uncertain outcome, why did this project even attempt random assignment? As envisioned by its original proponent, Mitchell (Mike) Sviridoff at the Ford Foundation, the Supported Work demonstration would assess whether a promising one-site program could be replicated in other locations and for different populations. Sviridoff envisioned a "respectable research component" and saw it as part of a try-small-before-you-spend-big vision of policymaking. But Sviridoff, who always thought big, had assembled a consortium of six federal funding partners and created an illustrious advisory committee, of which two members (Robert Solow and Robert Lampman) backed up by staff at HHS took the project in an unanticipated direction by insisting that "testing" meant using random assignment. When asked about it 35 years later, Solow attributed his determination to his training, saying "My first job was as a professor of statistics! I favored it because I wanted to have a defensible response." He and Lampman also shared the conviction that the research design had to be strong enough to detect what they anticipated would be, at best, small and complex effects (Gueron and Rolston 2013, 32, 483n13).

The result was a hybrid: Supported Work was both a demonstration and an experiment. As a demonstration, the project sought to provide sites with enough flexibility to create a realistic test of the administrative and other obstacles to replicating the multi-faceted program. As a social experiment, it needed sufficient standardization to define a "model" (the treatment), allow pooling data from multiple programs, and reduce the risk of evaluating a poorly implemented start-up period.

Why did 10 sites ultimately accept random assignment? As expected, initial opposition was strong. To do their jobs well, local staff had to believe they were helping people. Any intake procedure involves some form of rationing — first come first served, enrolling the more motivated first, allowing for caseworker discretion, or limiting recruitment so that no one is actually rejected. Staff overwhelmingly preferred those approaches to a random process in which they personally had to confront and turn away people they viewed as eligible and deserving. Yet for a social experiment to succeed, these staff had to be converted. They had to buy into the process or at least agree to cooperate fully with it. Otherwise, the study would be doomed, which is what many feared would happen to Supported Work. But the project did not fail. Relatively quickly, the process became familiar, complaints diminished, and random assignment was accepted. A high quality RCT was implemented, and the findings were not subject to the familiar methodological debate.

At the time, I and others attributed the ability to induce and discipline compliance to four conditions. The first and most important was money. Community organizations received millions of dollars to run a new and distinctive program conditional upon them playing by the rules, the most important of which was random assignment. There was also generous funding for research and data collection, including for in-person interviews to track 6,500 people for up to three years.

The second was strong nonfinancial incentives. The local Supported Work operators, referral agencies, and interest groups all viewed the program positively: it was voluntary; it offered paid jobs to underserved and hard-to-employ people at a time when others were advocating mandatory, unpaid work-for-your-benefits (workfare) programs; and there was an explicit commitment to high ethical and legal standards. Thus, the pitch used to recruit sites and train frontline staff stressed the rationale for and morality of random assignment. It was a specially funded demonstration that would provide enriched services that would not otherwise exist. It would not reduce service levels or deny people access to benefits to which they were entitled. It had the resources to enroll only a small number of those interested. It would increase services for one group without reducing them for another. Finally, though the program sounded like an idea that could not fail, there was as yet no evidence that it would actually help people. In these conditions, the demonstration's managers argued (1) a lottery was actually fairer than other ways to allocate scarce opportunities and (2) getting a reliable answer on effectiveness (and thus abiding by the study rules, including not helping controls) was consistent with the program operators' mission. Supported Work reaffirmed this message in its procedures, as it was the first social experiment to be covered by new federal regulations on the protection of human subjects. (At intake, through a process of informed consent, applicants were told about the lottery and the possible risks and advised of both the kind of data that would be collected in surveys — in some cases on illegal activities — and the strict procedures that would be put in place to protect confidentiality and limit data access.)

A third factor was the management structure and people involved. Given Supported Work's complexity, a new organization, MDRC, was created to impose tight central control on the project and balance operational and research priorities. MDRC, in turn, selected a team, which included people at Mathematica Policy Research and the University of Wisconsin's Institute for Research on Poverty who had played lead roles in the NIT experiments, to conduct the impact and benefit-cost analyses. This staffing decision was an early example of the continuity that persisted over the years, with later studies drawing, often directly, on the wisdom gained in earlier ones. Another force for continuity lay in MDRC's Board of Directors, of which one leading member, Robert Solow, served for a remarkable 40 plus years. Throughout his tenure, Solow was a consistent advocate for rigor and for the organization's pioneering use of random assignment in the evaluation of an expanding range of social and educational programs.

The fourth factor was the intentionally low profile. The location of random assignment in relatively small (several hundred volunteers per site) pilot programs run by community agencies gave the project a stealth quality that helped it fly below the potentially ruinous political and press radar.

In retrospect, Supported Work was an auspicious debut for using large-scale RCTs to evaluate operating programs. The incentives, commitment to ethical practices, and oversubscribed program won allies and gave MDRC clout to call the shots. The generous funding assured local interest and a large treatment-control treatment difference. The behind-the-scenes nature of the project averted controversy. Compared with what was to follow, it was a step out of the laboratory but not a movement into the real world of mainstream public agencies. From this experience, I and others concluded that the conditions that favored success were not just helpful but necessary for RCTs. Although it is probably true that, at the time, MDRC would not have succeeded without them (particularly the generous operating funds), subsequent events proved that these conditions were not indispensable.

In addition to demonstrating feasibility, the Supported Work findings (released in 1980) showed the value of using a control group to reach conclusions on effectiveness. Table 1 (which gives the percent of people in the treatment and control groups who were employed roughly two years after random assignment, as well as the difference or impact) points to three telling insights.[6]

Table 1 Percentage Employed Some Time Between the Nineteenth and Twenty-Seventh Month after Random Assignment: Supported Work Evaluation

Target group	Treatment group	Control group	Difference (Impact)
AFDC recipients	49.1	40.6	8.5**
Former addicts	56.5	53.0	3.5
Former offenders	56.5	53.3	3.2
Youth	62.6	62.6	0.0

Source: Gueron and Rolston 2013, 54.
**Statistically significant at the 5 percent level.

First, social programs can work, but not all prima facie good ideas do. Supported Work significantly increased the post-program employment of single mothers on AFDC and (not shown in Table 1) reduced their receipt of cash welfare. Given the prevailing skepticism, this success was heralded. But, the program did not have impacts on the three other groups.

Second, even for the AFDC group, impacts were modest. Although Supported Work boosted employment, the employment rate of the control group revealed that the big gain over the two years came from the economy and the myriad other factors that led people (almost all of whom were unemployed at the start of the study) to take a job.

[6] For more detail on the program and the findings, see Hollister, Kemper, and Maynard (1984) and MDRC Board of Directors (1980).

Third, high outcomes may not reflect high impacts. The demonstration's planners had expected that Supported Work would be least effective for AFDC women, since they had a harder time finding work, had competing child care responsibilities, and faced lower work incentives (they not only got jobs with lower wages but received welfare as an alternative source of income and their benefits would be cut if they worked). The data in column one appear to support this hunch: AFDC recipients were the least likely of the four groups to be working after participating in the program. However, evidence from the control groups disproves this expectation: the mostly male former addicts, offenders, and young school dropouts were also more likely to get jobs on their own, with the program making no significant difference. Thus, Supported Work succeeded with AFDC women not because the participants did so well (as measured by their outcomes) but because the corresponding control group members (without program aid) did so poorly. One implication was clear: traditional outcome-based performance measures (for example, how many enrollees were placed in jobs or left welfare) would have sent a false signal and led to wasted funds and less effective programs.

The magnitude, unpredictability, and complexity of the findings brought themes into focus that sharpened with time: (1) impacts, if they occur, are likely to be modest; (2) pay attention to the service differential, that is, do not focus only on the treatment group and the quality of the test program, but keep your eye on the control group (both their outcomes and the alternative services they and treatment group members receive); (3) beware of overreliance on outcome-based performance standards; and (4) look at impacts for key subgroups.

Supported Work also offered good news to people searching for ways to bring rigorous evidence to policy debates often dominated by claims made on a hunch or discredited on an anecdote. Once it became clear that the study had been meticulously implemented, there was widespread acceptance of the findings. The transparency of the method and the simplicity with which the results could be explained made random assignment a powerful communications tool. People differed on the implications for policy and questioned whether the impacts could be replicated on a larger scale, but there was not the familiar back and forth among experts that followed studies using more complex, and ultimately less interpretable, methods.

Nonetheless, even though Supported Work was a beautiful study that pioneered many methods used in subsequent RCTs, there was little pick up on the encouraging impacts for welfare mothers. We at MDRC attributed that to several factors: the project's origin (designed by elites with little state ownership), the nature of the program and findings (an expensive and complex model that produced gains similar to those later found for lower-cost approaches), and the 1980 election that ended federal interest. Although we had always known that positive results would not automatically lead to expansion and were chary about becoming advocates of the program rather than of the research, we went away thinking we had failed to build a constituency in the existing systems that would be waiting for the results and primed to act on them. Determined not to repeat that mistake, MDRC took a more inclusive and grassroots approach in subsequent experiments.

Social Experiments Reincarnated as a Partnership: Testing Feasibility Anew by Evaluating State Initiatives

What happened next was driven by long-term trends, the 1980 election, and institutional priorities. From the 1970s through the 1990s, welfare reform was a bitterly contentious political wedge issue stoked by increasing anger at a system that many felt encouraged dependency, undermined family structure, and unfairly supported people who could work but did not while others struggled in low-wage jobs. During these years, politicians ran for president or the state house on their record and claims as welfare reformers.

Several factors had spurred the erosion of support for AFDC as an open-ended entitlement. One was the dramatic growth in the rolls and costs. Created in 1935 as a program intended to support a small number of poor widows and wives of disabled workers (people not expected to work), it had swelled from 270,000 families in 1945 to 1,000,000 in 1965, 3,400,000 in 1975, 3,700,000 in 1985, and 4,900,000 in 1995 (Gueron and Rolston 2013, 481n3). A second was the change in who received welfare. The vast majority were not widows but divorced, separated, or never married women, reflecting what was widely perceived as an alarming dissolution of the family.[7] A third was that women across the country (including single parents with young children) were flooding into the labor force, often not by choice.

Together, 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 effects potentially included encouraging family breakup and teen pregnancy, discouraging women from earning a living, and making it easier for fathers to leave their families and avoid supporting their children. The result was that, over time, public debate shifted from whether mothers on welfare should work to who should work and how to make that happen, from voluntary programs such as Supported Work to mandates and obligations that would require people to work or participate in diverse work-directed activities, and later (in the 1990s) to whether there should be a limit on how long people could remain on the rolls.

Ronald Reagan's election in 1980 — following a campaign that capitalized on this hot-button issue — produced a dramatic change in welfare policy, the role of the states, and the nature and origin of research funds. The new administration saw workfare (work for your benefits) as the solution and, convinced of its benefits, was not interested in any rigorous evaluation. In Congress, however, there was no consensus on how to structure such a program or what different approaches might cost or yield. Consequently, rather than impose a nationwide vision, federal legislation in 1981 gave the states increased flexibility to undertake their own initiatives. At the same time, the administration, which viewed social science

[7] The proportion of children under 18 living with an unmarried mother had increased from 5 percent of white children (25 percent of black children) in 1965 to 15 percent of white children (50 percent of black children) by the early 1980s (McLanahan and Jencks 2015, 16).

researchers with suspicion and as advocates for the liberal policies they typically assessed, ended most funding for demonstrations and evaluations.

As a result, prospects for experiments looked bleak. The conditions that had nurtured Supported Work — generous funding, centralized clout, and an oversubscribed voluntary program — disappeared, in some cases permanently. More parochially, stunned by the cancelation of multiple studies (for an example, see Elmore 1985, 330) and having let go 45 percent of its staff, MDRC debated the chances and choices for survival. With a determination to maintain its focus on rigorous studies of programs for low-income people, MDRC dreamed up a partnership vision that proved to be the major turning point in the design of welfare experiments and within a decade both produced results of greater relevance and policy impact than the NIT or Supported Work experiments and became the model that flourished for the next 20 years.

With the specter of controversial state welfare reforms and no planned federal evaluation, MDRC sought Ford Foundation funding for an objective, outside assessment. The concept was to make a reality of Supreme Court Justice Brandeis's famous statement that the states were laboratories for experiments by taking the word "experiment" literally; that is, MDRC would convert into actual RCTs the initiatives that emerged as governors across the country responded enthusiastically to the opportunity to put their stamp on welfare.[8] Instead of one experiment that would test a centrally defined model in multiple sites (as in Supported Work), MDRC's resulting Work/Welfare Demonstration used RCTs to assess programs that reflected each state's particular values, resources, goals, and capabilities — but primarily required people to search for a job or work for their benefits — with random assignment integrated into the helter-skelter of normal agency operations (Gueron and Rolston 2013, 97-117).[9]

MDRC identified three key research questions to address in parallel studies in each state: Would the state run a mandatory program (and what would high participation and workfare look like in practice)? Would the reform reduce welfare or increase work and, if so, for whom? Would the change cost or save money? The nature of the programs and the absence of the key enablers of the Supported Work study drove a radically different vision for the

[8] To appreciate why governors played such a prominent role in welfare reform, it is important to understand how AFDC differed from some other programs. For example, in contrast to Social Security, which is fully funded by the federal government and operates under standard, nationwide rules, AFDC was designed as a federal-state partnership. On the one hand, the program was a federal entitlement, meaning that no person who satisfied the eligibility criteria and program rules could be denied benefits. On the other hand, it was a state program, insofar as the states retained substantial discretion over those rules and shared the cost with the federal government. Consequently, both states and the federal government had a strong financial incentive to reduce the rolls and, potentially, an appetite for reliable evidence on cost-effectiveness. Simultaneously, the unpopularity of the program created a political incentive for governors to compete for leadership as reformers.

[9] MDRC sought to place random assignment as early as feasible in the intake process (preferably at welfare application) because the reforms were expected to change the behavior not only of people who actually participated in the required activities but also of those who did not but were subject to the monitoring, the messaging, and the threat or reality of financial sanctions.

evaluation. Because the new mandates were intensely controversial, MDRC staff knew they would need the most rigorous evidence to defend any findings and thus chose random assignment. Because they anticipated at most modest impacts and had to assess each state initiative as a separate experiment, they knew they would need large samples, which ultimately involved 28,500 people. Because of the relatively limited research budget (the Ford Foundation's $3.6 million grant, which MDRC hoped to double, ultimately lasted more than five years), staff knew they could not track this vast sample using surveys but, for the first time in a large-scale RCT, would have to estimate impacts solely from existing administrative records.[10] This decision meant seeking reliable answers to the first order questions covered by these records and leaving the rest to future studies.

A social experiment had never before been attempted at this scale, in mainstream offices run by large bureaucracies, in mandatory programs, with no direct federal funds or role, with no special operating funds, and with no researcher leverage.[11] Further, MDRC would be testing relatively high-profile political initiatives that — although still viewed as demonstrations implemented in one or a few locations in a state — were hyped in gubernatorial and even presidential campaigns (one program was Governor Bill Clinton's initiative in Arkansas).

At a time when they were under pressure to launch new programs, why did some welfare commissioners accept the added work and potentially explosive risk of inserting a lottery into the stressful welfare intake process and participating in a demanding and independent study that could as easily shown failure as success? Not surprisingly, their initial reaction was disbelief. You want us to do what? Is this ethical? Will it impede operations? Will it explode?

In a courtship that extended over 30 states and lasted two years, MDRC gradually overcame these concerns, in eight states that met its requirements, by making specific design decisions, building relationships that nurtured trust, and marshalling five arguments to sell the project as a win-win opportunity.[12] As a group, these eight states were representative of both

[10] Albeit a decision of necessity, it had the advantage of limiting sample attrition and recall problems over the eventual five years of follow-up, although it raised some coverage issues, for example, by not tracking people after they left a state.

[11] It is useful to distinguish two aspects of social experiments that could be more or less subject to centralized control: the treatment being tested and the design and implementation of the research. On the former, the NITs were at one end of the continuum (total researcher control of the treatment), Supported Work a few steps along the continuum (a centrally defined model, with some room for local variation), and the Work/Welfare Demonstration at the other extreme (treatments defined by the states, with no researcher role). Along the research design control continuum, there was less variation. Researchers had full control of the design, random assignment process, data collection, analysis, and reporting in the NITs and Supported Work. In the Work/Welfare Demonstration, MDRC used the Ford Foundation funding to insist on a consistent research agenda and control of random assignment and data requirements, but also sought, in the partnership mode, to answer questions that were of interest to particular states.

[12] MDRC sought states that planned initiatives of sufficient scale to generate the needed samples, agreed to cooperate with research demands (not only random assignment but also monitoring and restricting services for a large share of the caseload), maintained and would share administrative records of sufficient quality, and could

nationwide responses to the 1981 law and the variety of local conditions (Gueron and Rolston 2013, 118-131).

The first selling point was the promise of a new style: a partnership that would answer *their* questions about *their* reforms, combined with a pitch on why getting answers required estimating impacts. The 1981 law's flexibility had put welfare commissioners on the spot. The system was unpopular and they were under pressure to get tough, but they understood the difficulty of implementing change and the diversity of people on the rolls. Although they had almost no reliable data on the likely cost and results of specific policies, at least some of them suspected that the job entry or case closure measures they typically touted would overstate success. The commissioners could grasp how the evidence from control groups in prior RCTs confirmed their doubts. But the challenge remained to explain why one needed an experiment, rather than some less intrusive design, to determine success, especially given the limited academic support and often outright opposition.[13] MDRC's response was fourfold: pretend there was a consensus and assume that welfare administrators would not follow or understand the econometric debate; educate them on the outcome-impact distinction and why outcomes would not answer their questions; expose the weaknesses of alternative designs; and offer a study that would accurately measure the real accomplishments of their programs, address other questions they cared about (for example, the impact on state budgets and insights on what may explain success or failure), and produce results that would be simple, credible, and defensible.

The second selling point was that random assignment was not some wacko scheme dreamed up by ivory-tower purists. It had been done before; it had not disrupted operations; and it had not blown up in the courts or in the press. The Supported Work experience got MDRC part way, but more powerful evidence for welfare administrators came from a small project, called the Work Incentive (WIN) Laboratories, that MDRC had managed in the late 1970s and that had lodged random assignment in a few local welfare-to-work program offices and thus involved civil servants facing normal pressures and performance requirements (Gueron and Rolston 2013, 66-87). However, the proposed state studies upped the ante: larger and much more political initiatives and the integration of random assignment into the high stakes welfare eligibility review process. To overcome these obstacles, MDRC promised to work with state staff and local community advocates to develop procedures that would be fair, ethical, and not overly burdensome; to provide extensive training so that frontline staff would

somehow provide 50 percent of the funds for the evaluation. This last condition proved by far the toughest, and most of the state contribution came from other sources. For a description of the programs and findings, see the individual state reports published by MDRC, Friedlander and Burtless 1995, Gueron and Pauly 1991, and Gueron and Rolston 2013.

[13] The opposition came from both qualitative and institutional researchers who said we were addressing narrow and relatively unimportant questions and economists who argued that statistical modeling could produce equally reliable answers at lower cost and that experimental results were likely to be biased (because random assignment altered the programs being analyzed) or of dubious scientific value (because they yielded no basic or cumulative knowledge). See Gueron and Rolston 2013, 270-272, 455-457.

understand the rationale for random assignment; and to produce results that would address pragmatic concerns.

The final three selling points were the offering of a subsidized study that met the then vague but useful federal requirement for an independent assessment of the waivers to welfare rules, which most states needed to implement their initiatives;[14] modest assistance on program design; and prestige from selection for a high-profile Ford Foundation initiative (although at the time no one remotely anticipated the visibility that would come to participating states).

Nonetheless, enlisting states was a tough sell. There was always pressure to use weaker, less intrusive research designs. That the pitch ultimately worked is why I have called the welfare commissioners the heroes of the survival and reincarnation of welfare experiments. Their unflinching support once they had signed on was the major reason why random assignment was the dog that did not bark and why no state dropped out of or sought to undermine the studies, despite the relentless beating some of them took from having their programs assessed using the new and tough metric (impacts) and at a time when governors in other states trumpeted their success and built their reputations based on misleading but numerically vastly higher outcomes (Gueron and Rolston 2013, 128-31, 256). In an effort to assist participating states and debunk these claims, MDRC repeatedly sought to educate the press, advocacy groups, congressional staff, and senior state and federal policymakers about the erroneous use of and unrealistic expectations generated by hyping outcome data and the truth of the more modest results from the RCTs.

Collaboration and partnership are often empty slogans masking business as usual. However, in 1982 it was clear that MDRC's powerlessness vis-à-vis the states required a genuinely new style, in which leverage was based not on holding the purse strings, and as a result calling the shots, but on the quality of working relationships, the usefulness of the findings, and the creation of a strong mutual interest in and commitment to obtaining credible answers. The outcome was positive: by trading control and uniformity of the operating programs for relevance and ownership, the states had a greater commitment to the treatments and ultimately the RCTs, which in turn provided a built-in constituency for the results (Gueron and Rolston 2013, 105; Blum with Blank 1990).

The partnership model also had the unanticipated benefit of treatment replication. Six of the initial states (plus a second RCT in San Diego that HHS initiated and MDRC conducted) sought to implement variations on the theme of work requirements, with job search as the first and major activity followed (for some) by unpaid work experience. But in the context of welfare experiments, replication did not mean reproducing an identical, centrally specified model. Just as welfare benefit levels differed greatly across the country, so did the reforms' specific design, targeting, goals, cost, and implementation (the messaging, participation rates, and intensity

[14] The subsidy came mostly from the Ford Foundation and, indirectly, federal special demonstration and matching funds. For a discussion on the critical role of the 50 percent uncapped federal match for state evaluations under AFDC, see Gueron and Rolston 2013, 134, 258-259, 386.

and nature of services). (For example, the nature of job search varied from individual job searches, in which people were expected, on their own, to follow up, and report back on job leads, to job clubs, in which program staff might provide instruction on resume preparation and interviewing, a phone room for contacting prospective employers, and job leads.) They also differed in context: urban or rural, labor market conditions, and the extent of alternative services available to people in the treatment and control groups. Since each state program was a separate RCT, this created a form of replication that, as discussed below, greatly increased the influence of the findings, as it became clear that most of the reforms had impacts in the desired direction.

However, the shift in authority (the studies were conducted under state contracts), combined with the mandatory nature of the initiatives and the commitment to providing useful findings, prompted a controversial departure from past RCTs. Because states insisted on learning the effect of their reforms on the full range of people required to participate (not just those who might volunteer to be in the study or the program, if given a choice), eligible people could not opt out of the program, or of random assignment, or of any follow-up that relied on the states' own administrative records. This stipulation assured generalizability of the results to the universe of people subject to the new requirements and made the studies more akin to natural field experiments.[15]

The strategy of state-based experiments was a success. Random assignment worked, as did the reliance on administrative records (Gueron and Rolston 2013, 185-90). More importantly for the participating states, the findings were judged encouraging. States had sought, to varying degrees, to make progress on four goals: increase employment, reduce dependence on public assistance, save money, and make families better off. Although not articulated as such, they likely also shared a fifth goal, cost-effectiveness, defined as the impact per dollar spent or the "bang for the buck." (This last goal is particularly relevant to welfare reform initiatives, since they are, by intent, mass interventions that seek both to change individual behavior and to reduce the welfare rolls.[16])

[15] At each site, random assignment was used to create a treatment group that was subject to the new program and its requirements and a control group that was excused from both the newly required services and the threatened financial penalties for noncooperation. People in both groups would be told they were in the study and subject to a lottery, informed of the grievance procedures, and given a choice about responding to any special surveys. Most welfare advocates did not object to the elimination of a general informed consent because at the time they viewed the new mandates as punitive and were glad that the control group was excused from potential sanctions (Gueron and Rolston 2013, 186-188). For a discussion of the level of control in laboratory experiments (where people are aware of their participation and give informed consent) versus natural field experiments (where people are assigned covertly, without their consent), see Al-Ubaydli and List (2014).

[16] The cost-effectiveness of social programs always matters, but it is made particularly salient by a fundamental difference between the 1980s welfare reforms and programs such as Supported Work. Most of the state initiatives were viewed as a dry run in a few locations of potential statewide (or even nationwide) reforms. The evaluations were designed explicitly to assess the impact of changing the service delivery system — including mandatory administrative practices, case management, and multiple components — for all eligible people in the demonstration areas. In contrast, the Supported Work evaluation assessed a single activity intended to reach a fixed number of volunteers. The more cost-effective a state's initiative, the greater is its ability to reach a larger share of the caseload within a given budget and hence to produce a bigger aggregate or total impact. This

The findings, released between 1984 and 1988, showed progress on most fronts. The programs generally increased average employment rates and earnings and, somewhat less consistently, reduced welfare receipt. Surprisingly, most of them also saved money, generating cumulative reductions in AFDC payments and other transfers that within a few years exceeded the programs' net costs. The combination of modest impacts on behavior and low costs also made most of them highly cost-effective. There were, however, minimal or no impacts on family income or poverty (Friedlander and Burtless 1995, 32, 87-101; Gueron 1990; Gueron and Pauly 1991, 142-154; Gueron and Rolston 2013, 182-185).

An in-depth analysis of the four programs that had five years of follow-up showed that average impacts (ranging from 3 to 7 percentage point increases in quarterly employment rates and 0 to 8 percentage point reductions in the monthly rate of AFDC receipt) remained strong for three to four years, after which the controls began to catch up with the treatment group.[17] The study concluded that the programs encouraged more people to start working and to leave welfare sooner than they would have without the reforms, but generally did not help them get higher-paying or more stable jobs (leaving many with little income and back on the rolls) and did little to reduce welfare for more disadvantaged, potential long-term recipients (Friedlander and Burtless 1995, 2-3, 16, 88-101).

Using RCTs to Test Full-Scale Programs: the Fight Got Tougher

By 1986, the terrain for welfare experiments had changed. MDRC had shown that RCTs testing state initiatives were feasible. A number of senior people in the Reagan administration had become strong supporters. Some governors and commissioners had seen firsthand that such studies not only were not toxic but also could contribute to their claim for leadership as welfare reformers and produce valuable lessons that brought them unanticipated renown.

What followed over the next 15 years made the welfare saga exceptional: a flowering of RCTs that has been called the "golden age of social welfare experimentation" (Manzi 2012, 184). Separately and in interaction, MDRC, other research firms, HHS, and state administrators built a coherent body of evidence about the effectiveness of the major policy alternatives. After identifying what it considered the key policy options, MDRC sought to assemble clusters of places planning or willing to try out those approaches, aiming to repeat its early 1980s strategy of turning the dynamic state reform context into an opportunity to learn (at times, by again

consideration (which highlights the importance of cumulative welfare savings) is fundamental when comparing results for higher- and lower-cost approaches (Gueron and Rolston 2013, 103, 207-208, 425; Gueron and Pauly 1991, 70-78; Friedlander and Burtless 1995, 71; and Friedlander and Gueron 1992).

[17] Friedlander and Burtless (1995, 8-9, 58-60) caution readers to view this "catch-up" as a lower-bound estimate of the long-term effects of permanent programs. Because the original evaluation plans had envisioned a relatively short follow-up, the embargo on enrolling controls in the test programs lasted for only two years following random assignment, although the programs generally continued after that time. As a result, the five-year follow-up included years during which some people in the control group might have been subject to the mandates and services, possibly reducing the late-year estimates and making impacts appear less long lasting.

leveraging Ford Foundation grants). Staff at HHS led by Howard Rolston in what was then the Family Support Administration and Michael Fishman and others in the Office of the Assistant Secretary for Planning and Evaluation launched increasingly ambitious experiments, culminating in the largest and most complex welfare RCT, and embarked on a five-year journey with the U.S. Office of Management and Budget (OMB) to require states that sought waivers in order to modify standard policy to assess their initiatives using a control group created through random assignment. In 1992, after ups and downs, an RCT became the required yardstick by which to measure the fiscal neutrality of the explosion of waivers that states requested in a push for more — and more ambitious — reforms.[18]

The result was an accretive agenda that looked carefully orchestrated but in reality emerged from a feedback loop, in which experiments generated findings and raised substantive and methodological questions and hypotheses that prompted successive tests. (See Table 2 for examples.)

Table 2 Evolution of Welfare Research Agenda

Findings from prior studies	**Prompted new questions and tests**
The early 1980s low-cost mandatory job search/workfare programs produced small-to-modest increases in employment and (less consistently) reductions in welfare for single mothers with school-aged children	Would remediation of basic education deficits increase success, particularly for more disadvantaged recipients?
	Would similar approaches succeed with mothers of younger children? With teen parents?
	Would work-related mandates help or hurt young children in welfare families?
	Would impacts increase if ongoing participation was required as long as people remained on welfare?
Single- or multi-county demonstrations and pilot programs produced encouraging results	Could success be replicated or improved upon in full-scale, statewide programs?
Programs requiring some combination of job search, workfare, and basic education increased work but did little to reduce poverty	Would programs that supplemented earnings increase work, reduce poverty, and benefit children?
	Would extending services or mandates to the noncustodial fathers of children on welfare increase child support payments or improve outcomes for children?
Comparisons of impacts across sites	Could this be confirmed in multi-arm RCTs

[18] Since 1962, HHS had had the authority to grant states waivers of AFDC program requirements in order to try out innovations. But only after 1992, and thus after the most influential of the welfare experiments, was a quid pro quo firmly implemented, in which states could not get waivers without conducting an RCT. The logic was straightforward: HHS and OMB had learned that RCTs were feasible and much more reliable than alternative research designs. As a result, and in order to assure that waivers did not become an intended or unintended drain on federal budgets, they insisted that they be used to measure fiscal impact and to allocate costs between federal and state budgets (Gueron and Rolston 2013, 156-159, 217-261).

suggested certain approaches were more effective than others	testing varied approaches in the same sites?

Source: Author's compilation.

The initial effect of this expanding agenda was that a tough fight got tougher. The strongest opposition arose after senior officials in California and Florida, in late 1985 and 1989, invited MDRC to conduct random assignment evaluations of their respective statewide programs: Greater Avenues for Independence (GAIN) and Project Independence (PI). The officials' reasons differed, but neither state was driven by the need for waivers. In California, some people in the legislature and state agencies had seen firsthand the problem-free implementation of MDRC's earlier RCTs in San Diego and the usefulness and influence of the findings. As a result, once they agreed that GAIN had to be rigorously evaluated, they quickly concurred across party lines that rigor meant random assignment.

In Florida, after the agency charged by the legislature to determine effectiveness had been attacked for producing conflicting findings from successive studies using nonexperimental methods, Don Winstead, the key state official, sought guidance from Robinson Hollister, the chair of a recent National Academy of Sciences panel, who advised him to do it the right way and use random assignment. In contrast to the situation in California, Winstead had no familiarity with RCTs but, after reading reports from earlier experiments and Senator Daniel Patrick Moynihan's statements about the role of such research in the 1988 federal legislation, was persuaded that "the only way to get out of the pickle of these dueling unprovable things. . .and salvage the credibility of the program. . .was to get an evaluation of unquestioned quality and go forward" (Gueron and Rolston 2013, 301).[19]

Yet, despite strong support at the top and for the first time having random assignment written into legislation in California and Florida, what followed were legal and ethical objections that went way beyond those raised in the first generation of state studies. In Florida, a firestorm of opposition flared up that almost led the legislature to ban control groups and in the process would have both jeopardized a major federal research project and potentially poisoned the well for future studies (Gueron and Rolston 2013, 281-287, 298-309).

What explains the fierce reaction? The California GAIN and Florida PI programs were not just more of the same. They were more ambitious in scale, permanence, and prominence, and they also shifted the balance between opportunity and obligation. Earlier experiments had assessed reforms designed by researchers or funders (such as Supported Work and the NITs) or state-run initiatives that though large compared with prior evaluations were implemented on a trial basis in a few locations. Now, for the first time, random assignment was proposed to evaluate two programs that were intended to be universal (covering all who met the mandatory criteria), full scale, ongoing, and statewide. Further, the numbers were huge: GAIN

[19] Critically important, Winstead had strong support from the Secretary of Florida's Department of Health and Rehabilitative Services, Gregory Coler, who, notwithstanding negative findings from MDRC's earlier random assignment evaluation of the program he had run in Illinois, sought out such a study when he took over in Florida, having seen firsthand the credibility that Congressional staff and the press accorded to findings from experiments.

was the largest and most ambitious welfare-to-work program in the nation, with a projected budget of over $300 million a year and targeting 200,000 people, 35,000 of whom were ultimately subject to random assignment (Gueron and Rolston 2013, 276-278).

As a result, both evaluations raised an ethical red flag: Would the creation of a control group reduce the number of people served? Would it in effect deny people access to a quasi or real entitlement? The specific activities added another element. In earlier RCTs of mandatory programs, most welfare advocates had not objected to excluding controls from the services and penalties, in part because the programs were viewed primarily as imposing burdens not offering opportunities. Now, when the required activities included remedial education, denial of service became more controversial.

In combination, these differences meant that, far from being stealth evaluations, they appeared immediately and vividly on the political and press radars. In California, MDRC staff were called Nazis and a senior legislator who believed deeply in the value of education threatened to close down the study. In Florida, a lethal combination of gubernatorial politics, a concerned legislator, and ill will between the advocacy community and the welfare agency fed an explosion of inflammatory press. Headlines accused the state and MDRC of treating welfare recipients like guinea pigs and implementing practices that were shameful, inhuman, and akin to those used in the infamous Tuskegee syphilis study. Even in this pre-Internet era, the flare-up ricocheted to newspapers across the country, threatening other HHS experiments.

Proponents in the two states, MDRC, and HHS ultimately prevailed (showing the fallacy of claims that random assignment can be used only to assess small-scale operations) by both drawing on know-how gained in the earlier state studies and leveraging new forces. The first, and most important, was the unflinching stand taken by California and Florida officials who did not walk away when attacked, despite withering criticism. No researcher or research firm could have overcome this level of opposition alone. The determination of state officials to get an independent and credible evaluation — one that would address their questions but that they were well aware could expose their failure — was inspiring. Thus, when threatened with lawsuits, Carl Williams, California's GAIN administrator, said he was simply unwilling to supervise a program of that size and complexity unless it had a really sound evaluation, declaring, "We were going to get random assignment one way or another." When asked why he fought for the study, Winstead replied, "It sounds sort of naïve, but I became convinced that it was the right thing to do. . . . If we're going to put thousands of people through something, we ought to be willing to find out whether or not it works" (Gueron and Rolston 2013, 281, 285, 307).

The second new factor was the slow shift in academic backing for random assignment, reflected in and prodded by two events. The first, in 1985, was the release of authoritative reports from the National Academy of Sciences and the Department of Labor, publications that MDRC cited over and over again to encourage allies and convert opponents (Betsey, Hollister, and Papageorgiou and Job Training Longitudinal Survey Research Advisory Panel). Both expert panels concluded that they did not believe the results of most comparison group studies —

including the Department of Labor's $50 million or so outlay on conflicting econometric evaluations of the nation's major job training program — and saw no alternative to random assignment given existing statistical techniques if one wanted to produce credible data on effectiveness. The second event was an unexpected legacy from Supported Work. Not only did the demonstration show that it was feasible to use a field experiment to evaluate a large-scale employment program, but it also provided a public use file that, for the first time, offered an intriguing way to find out if alternatives could have done as well. Robert LaLonde's groundbreaking study, published in the *American Economic Review* in 1986, did just that by testing whether econometric estimates — using eight carefully constructed comparison groups drawn from the Current Population Survey and the Panel Study of Income Dynamics — could reliably reproduce the experimental findings. His negative conclusion had a profound influence.[20]

The third factor was the successful effort to build and then mobilize a community of converts and fans (including advocates, public officials, funders, academics, practitioners, and state and federal legislative, congressional, and agency staff) who recognized and valued the distinctive quality of the evidence from RCTs and became allies in defending the studies and their results. This factor became particularly important when MDRC and state staff in Florida, fearful that a successful lawsuit or ban on control group research in the state risked widespread contagion, used endorsements from these sources and one-on-one meetings with dozens of legislators to sell the merits and ethics of RCTs.

The final and most decisive factor in both states was a budget shortfall. Despite the rhetoric of universal mandates, the reality was that there were not enough funds to reach everyone. Once it became clear that services would have to be rationed and some eligible people denied access (but not as a result of the study), a lottery struck the objecting legislators as a fair way to give everyone an equal chance. (The California and Florida experience also led HHS to prohibit using RCTs to test entitlements.)

The findings from the GAIN evaluation addressed a number of the issues raised by the early 1980s state studies (see Table 2). The GAIN approach reflected the hope that emphasizing basic education for those with limited academic skills and helping the rest get a job quickly would produce better results (particularly for long-term welfare recipients) than would the shorter-term, primarily job search programs, and that the higher cost would be worth it. The effects for the six study counties combined were mixed. GAIN outperformed the earlier programs on some measures, generating larger, longer-lasting impacts (still robust five years later) and having greater success with more disadvantaged recipients. Nonetheless, the big picture remained in the range of modest but positive: in the average three months during the five years of follow up, 28 percent of single mothers assigned to the program worked,

[20] LaLonde (1986, 604) states: "This comparison shows that many of the econometric procedures do not replicate the experimentally determined results, and it suggests that researchers should be aware of the potential for specification errors in other nonexperimental evaluations." Subsequently, Fraker and Maynard (1987) also used the Supported Work data and reached similar conclusions. Bloom, Michalopoulos, and Hill (2005) describe the numerous studies that followed, drawing on data from other experiments.

compared with 24 percent of control group members; by the last three months, 39 percent received some AFDC benefits compared with 42 percent of control group members. Further, in contrast to most of the earlier state studies, GAIN did not pay for itself, returning $0.76 in budget savings for every public dollar spent to run it (with net costs calculated as the difference in the average cost of all services received by program and control group members). On the other hand, GAIN did better in raising participants' income, leading to a 3 percentage point reduction in the share of families in poverty (Gueron and Rolston 2013, 287-289; Freedman et al. 1996).

The findings for one county, Riverside, however, were strikingly more positive, with better results overall and among low-skilled recipients, lower costs, and a highly cost-effective program that returned to taxpayers almost three dollars for every dollar invested. For the first time, a welfare-to-work program produced effects that broke out of the modest range. These findings raised an obvious question: What explained Riverside's success?

GAIN had given counties substantial discretion in how they implemented the program. Although Riverside provided a mix of activities (and had other special features), it emphasized getting a job quickly and, for those deemed to need basic education, offered work-focused short-term education or training. In the early years of the study, welfare directors in counties that had made a greater investment in education argued that it would pay off in the long term, particularly for people without skills or a high school degree. But the two-, three-, and five-year results confirmed a different story: Though impacts in the other counties grew over time, Riverside stayed in the lead on most measures and, crucially, proved to be the most successful with more disadvantaged recipients (Riccio, Friedlander, and Freedman 1994; Freedman et al. 1996; Gueron and Rolston 2013, 289-290; Gueron 1996; Gueron and Hamilton 2002).

What Works Best? A Multi-Arm Test of Labor Force Attachment Versus Human Capital Development

This counterintuitive finding along a major liberal-conservative fault line — work first versus education first — attracted attention in Washington and across the country (de Parle 2004, 111). However, since it came from comparing RCT results across California counties that differed not only in their program designs but also in labor market conditions, alternative services, and welfare populations — a nonexperimental comparison — it cried out for more rigorous confirmation. Hotz, Imbens, and Klerman (2006) sought to do this in a study that extended the GAIN follow-up to nine years and controlled statistically for county differences in pre- and post-program background and local conditions. They concluded that the other counties eventually caught up with and then surpassed Riverside's employment and earnings impacts and called for a reconsideration of the value of "training components that stress the development of work-related skills."[21]

[21] Hotz, Imbens, and Klerman did not address the relative success of the counties in meeting GAIN's other goals, including reducing cumulative welfare outlays and increasing cost-effectiveness. (See footnote 16.)

Although an important extension to the GAIN evaluation, the Hotz, Imbens and Klerman conclusion was still based on a nonexperimental analysis in six counties in one state. It raised a challenge: Was it possible to get a more definitive, experimental answer to this key policy question? Fortunately, a response was already in the works. In mid-1989, HHS had launched and MDRC was selected to conduct the most ambitious of the welfare experiments: randomly assigning 57,000 individuals to 11 programs at seven sites to evaluate the Job Opportunities and Basic Skills Training (JOBS) program, the major component of the 1988 federal welfare legislation. JOBS extended the requirement for participation in work-directed activities to mothers with younger children and emphasized serving people at risk of long-term dependency (Hamilton et al. 2001; Gueron and Rolston 2013, 311-352). The major hypothesis underlying JOBS (as with GAIN) was that providing remedial education to people with low basic skills was the strategy of choice for helping them to get better and more stable jobs, increase their family's income, and reduce their return to the rolls. It was expected that programs emphasizing education and training would be longer and more costly. The central questions, as in GAIN, included whether they would produce greater or longer-lasting impacts and be cost-effective in budgetary or other terms.

The centerpiece of the JOBS evaluation was an innovative and daring head-to-head test at three sites in which welfare recipients were randomly assigned either to a no-JOBS control group or to one of two different approaches: mandatory job-search-first programs, called labor force attachment (LFA) programs, that encouraged people to find employment quickly, or mandatory education-or-training-first programs, called human capital development (HCD) programs, that emphasized longer-term skill-building activities, primarily basic or remedial education, GED preparation, and, to a lesser extent, vocational training (but not college).[22] In contrast to the GAIN evaluation, this three-group design could produce *experimental* estimates of not only the impacts of each of the strategies (LFA versus a control group and HCD versus a control group) but also their differential effectiveness (LFA versus HCD). Overcoming MDRC's and HHS' initial concerns about feasibility, the two treatments and the multi-arm research design were successfully implemented at three very different sites: Grand Rapids, Michigan; Riverside, California; and Atlanta, Georgia (Hamilton et al. 1997).[23]

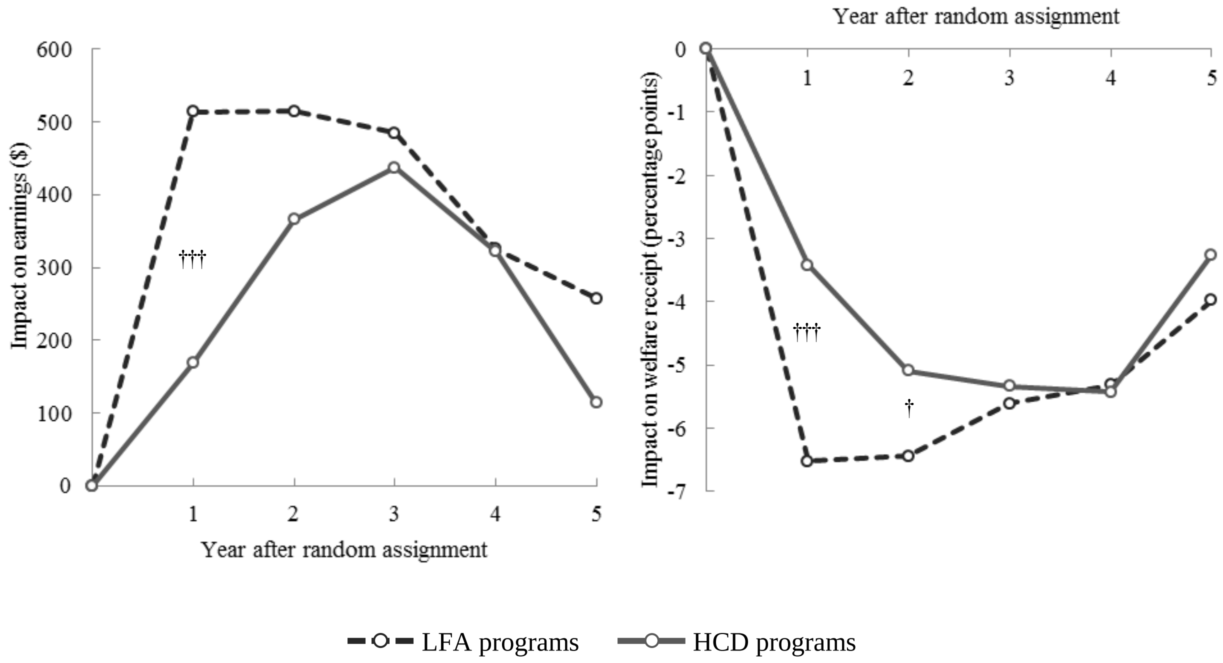
Figure 1 shows the impacts (the difference between averages for the treatment and control groups) on single mothers' earnings and welfare receipt of the LFA and HCD programs

[22] The LFA strategy reflected the view that the best way to build work habits and skills was by working, even at low wages; the HCD strategy was based on the belief that education and training should come first so that people could gain the skills required for them to get better jobs. Although both approaches included elements of the other (for example, people in the LFA programs who did not find work through job clubs could be assigned to short-term education or training or to unpaid work and people in the HCD stream could later be assigned to job clubs), they conveyed different messages and emphasized different activities (Hamilton et al. 1997).

[23] Although MDRC had employed multi-arm designs in three welfare RCTs during the early 1980s, these RCTs had tested whether adding workfare after job search increased effectiveness. The JOBS evaluation was much more ambitious. It required welfare agencies to operate two distinct comprehensive programs simultaneously. In addition, the JOBS evaluation used multi-arm designs to assess alternative case management strategies and to determine the separate effect of the program's services and its participation mandate (Gueron and Pauly 1991, 164n37; Gueron and Rolston 2013, 322-338).

at the three sites combined for each of the five years following random assignment. Both approaches increased earnings and reduced welfare, but the time trends differed. The LFA programs moved people into jobs and off welfare more quickly and thus had larger short-term impacts (that is, the LFA and HCD impacts differed significantly from each other in the first year or two, depending on the outcome measure). However, by the third year, the HCD programs had caught up: The gap between the two lines narrowed and was no longer statistically significant (Hamilton, 2002, 32). But Hotz, Imbens, and Klerman had suggested that five years would not do justice to the HCD strategy. MDRC's final report on the JOBS evaluation concluded that the five-year trends made it unlikely that the story would change, if longer follow-up were available. This conclusion was confirmed when Freedman and Smith (2008a, 2008b) tracked impacts up to 15 years after random assignment.

Figure 1 Impact on Earnings and Welfare Receipt, by Approach and Year: JOBS Evaluation



Source: Hamilton 2002, Figures 8 and 9.
Notes: The impacts shown are averages for sample members in the three LFA and HCD programs.
Daggers (†) denote statistical significance levels for LFA-HCD differences: † = 10 percent; †† = 5 percent; ††† = 1 percent.

Since the HCD programs did not surpass the LFA programs in the out years, the earnings gains and welfare savings over the entire five-year period (for both the three sites combined and in each site) were either the same for the two approaches or larger for the LFA programs. Furthermore, because the HCD programs were from one-third more expensive to nearly twice as expensive as the LFA programs that operated in the same site, the LFA programs proved to be much more cost-effective. As a result, for the same cost, the programs could reach more people and have larger aggregate impacts.[24] Particularly disappointing for advocates of the HCD approach, these findings held true both for program enrollees who lacked a high school diploma or a General Educational Development (GED) certificate — the subgroup of welfare recipients who were expected to derive the greatest benefit from the initial investment in basic education — as well as for those who already possessed one of those credentials (Hamilton 2002; Hamilton et al. 2001; Gueron and Rolston 2013).

The JOBS evaluation demonstrated the value of the multi-arm, multi-site designs by providing convincing evidence that, in mandatory programs for welfare mothers, the rigid job-search-first approach was more successful than the rigid education-or-training-first approach. However, this finding did not mean that there should be no role for education or training in welfare-to-work programs. (A cross-site comparison of results from 20 welfare RCTs suggests that the most successful ones used a mixed strategy, in which some people were urged to get a job quickly and others required to enroll in work-focused, short-term education or training [Gueron and Hamilton 2002].) It also did not imply that other types of training or postsecondary education or programs targeting different populations or volunteers would not be more effective. (For example, see Card, Kluve, and Weber 2015 and Hamilton 2012.)

The Momentum Shifts

By the early 1990s, four changes had shifted the momentum further in favor of random assignment: the evidence of the feasibility and payoff from the more ambitious and sophisticated tests, the visibility of the completed experiments and participating states, the final success of the HHS/OMB effort to make random assignment the quid pro quo for waivers, and the slowly gathering support among academics.

The result was that, instead of researchers or funders having to sell RCTs to states that accepted them reluctantly as the new price for HHS waivers, the reverse sometimes occurred. This outcome was most notable when the Canadian government, the New Hope program in Milwaukee, and the state of Minnesota proposed reforms to make work pay more than welfare by supplementing people's earnings if (in most cases) they worked full time. All three sought random assignment evaluations as the way to convince a wider audience of the value of their approaches. For them, despite the challenges (particularly in implementing multi-arm designs to determine which aspects of complex programs drove any impacts), experiments had been transformed from high-risk endeavors to pathways to recognition. Because MDRC used consistent measures and research designs in these and many of its earlier state studies, it was relatively easy to compare the success of different strategies — for example, mandatory welfare-to-work programs and earnings supplements — in advancing reformers' diverse goals.

[24] Advocates of the LFA and HCD strategies had anticipated that administrators might face a tradeoff in advancing different goals, with LFA more successful in saving money and HCD more successful in reducing poverty. However, the HCD programs did not have larger impacts on either outcome. This reflects the finding that, contrary to their goal, the HCD programs did not produce more earnings growth or increase the likelihood of sample members' getting more stable or higher paying jobs. They also did not differentially affect the well-being of sample members' children.

As shown in Table 3, comparing the two strategies revealed that mandatory welfare-to-work programs did better in reducing welfare dependency and government spending, earning supplements did better in reducing poverty and benefiting young children, and no single policy maximized all goals.[25]

Table 3 Trade-Off in Benefits from Two Reform Strategies

Goal	Earnings supplement	Welfare-to-work mandate
Reduce poverty	Yes	Usually no; at best, small
Benefit young children	Yes	No
Increase work	Usually yes	Yes
Save money	No	Often yes, but depends on design
Reduce welfare	Depends on design	Yes

Source: Gueron and Rolston 2013, 385.

Starting in 1996, when AFDC was replaced by a block grant to states, the incentive structure for RCTs shifted again. States could now redesign welfare on their own (no federal waivers needed) but could not tap federal matching funds for evaluation.[26] Fortunately, HHS's commitment to RCTs did not change. After a few years, during which it focused on sustaining the most valuable waiver experiments, HHS shifted gears and took the lead in launching multi-site experiments that addressed questions of interest to states in the new TANF environment. By the early 2000s — signaled in part by the creation of the Institute of Education Sciences in the U.S. Department of Education in 2002 — the explosion of interest in experiments was in full swing (Gueron and Rolston 2013, 388-422, 455-471).

Useful and Used

As stated above, the architects of the welfare experiments sought not only to obtain reliable evidence of effectiveness but to make the studies useful and to increase the potential that they would be used. A number of people close to the transformation of the U.S. welfare system — both the radical 1996 law that ended the AFDC entitlement and imposed tough work requirements and the 1988 bill that required participation in activities designed to enhance employability — have suggested that the experiments were unusually influential in shaping attitudes, legislation, and practice. For example, Ron Haskins, head of the Republican staff on the welfare subcommittee of House Ways and Means during these years, stated:

> Work really is the central issue of welfare reform, and the idea of work took on even more significance as an important, achievable goal because of the experi-

[25] For more information about the treatments, findings, and tradeoffs, see Bloom and Michalopoulos (2001), Gueron and Rolston (2013), Berlin (2000), Greenberg, Deitch, and Hamilton (2010), Morris et al. (2001), and Morris, Gennetian, and Duncan (2005).

[26] Thus, the block grant structure increased states' financial incentive to reduce the rolls, since they would reap all of the savings if people left welfare and bear all the costs for any expansion, but reduced the incentive for evaluation.

> ments. They took that potentially contentious issue off the table by showing that states could actually do things to get more people to work and save money. As a result of the experiments, by something like osmosis everybody in Congress came to understand this, it became the new conventional wisdom, and this had a dramatic effect on the welfare debate. . . . It is the best story I know of how research influenced policy (Gueron and Rolston 2013, 296-297).[27]

None of these people claimed that the legislation tracked the experiments (central parts of both bills reflected hunches that went way beyond the findings) or that politics, philosophy, and values were not much more important, but they did offer four reasons why this group of studies had an outsized effect (Gueron and Rolston 2013, 190-216, 436-443).

The credibility of random assignment, replication, and relevance

A major rationale for RCTs was the belief that policymakers could distinguish — and might privilege — the uncommon quality of the evidence. For a number of reasons, it seems that this hypothesis was often the case: the simplicity and transparency of the method; the growing consensus in the research community that alternative designs would fall short; the indication that performance measures such as job placements overstated success; and the replication of results in diverse conditions and in small, medium, and full scale programs.[28] All of these contributed to a bipartisan agreement that the RCTs offered an unusually reliable and objective yardstick.

The reaction to the studies suggested that policymakers valued external validity, though not in any formal statistical sense. The strategy described earlier — judgmentally selecting states which were representative along the dimensions that politically savvy folks viewed as likely to affect success (for example, strong and weak labor markets and administrative capacity), conducting experiments in ordinary offices, and having samples that were unscreened and large enough to produce valid estimates for each location — provided convincing face validity that the findings could be generalized beyond the study sites. As an example, Erica Baum (recruited by Senator Moynihan to draft the Senate's version of the 1988 legislation) points to the importance of finding positive results across nearly all the states studied, despite the variation in design, conditions, cost, population, attitudes, and administrative capacity. She particularly highlighted that regular staff in regular offices delivered the programs:

> This is no minor matter. In the past, elaborate programs pilot-tested by sophisticated social scientists or a small number of program experts produced worthwhile findings. But when the programs were transplanted to real-world social

[27] For different views on how and why these studies did or did not influence policy and practice, see Gueron and Rolston 2013, 190-215, 292-298; Baron 2013; Baum 1991; Greenberg, Linksz, and Mandel 2003; Haskins 1991; Rogers-Dillon 2004, 46; Szanton 1991; Weaver 2000.

[28] This high rate of replication contrasts with low rates in other fields (Manzi 2012) and what Begley and Ioannidis (2015) call the "reproducibility crisis" in biomedical research.

> agencies . . . the positive results disappeared. Since MDRC found that diverse state and local administrators could succeed on their own . . . we could be relatively confident that . . . other cities, counties, and states could do likewise (Gueron and Rolston 2013, 195).

The findings from comprehensive studies

The experiments had been structured strategically to test the major reform options and address the key concerns of liberals and conservatives. Although the effectiveness findings were the centerpiece (and the focus of this chapter), they were by no means the only evidence that the designers had thought would be important. Random assignment was always viewed as the skeleton on which to build studies using multiple techniques to answer a range of questions about program implementation and the factors that made programs more or less effective. The reaction showed that varied parts of the research did indeed matter to different audiences.

The fact that the impacts from mandatory welfare-to-work programs were relatively consistent and in the desired direction (increased work, reduced welfare) was critical. However, their absolute magnitude also mattered and played out differently in 1988 and 1996. In the early period, the modest gains prompted expanded funding for work programs; ten years later, limited success in the face of an increase in the rolls and the more stridently partisan context convinced some policymakers that a kind of shock therapy was called for.[29]

The findings on participation rates, suggesting that states could be trusted to impose serious obligations, contributed to the push for block grants. The finding that, under certain conditions, welfare recipients considered workfare fair changed the views of some originally hostile to requiring unpaid work. The counterintuitive evidence that programs emphasizing rapid employment had larger impacts than those requiring basic education contributed to a transformation of state programs. And the benefit-cost lesson — that up-front outlays were sometimes more than offset by rapid savings from reduced transfer payments and increased taxes as people went to work — provided unanticipated confirmation that social programs could be worthwhile investments and affected the all-important Congressional Budget Office estimates of the cost of legislative proposals (Gueron and Rolston 2013, 173).

The timeliness of results

The timing of results also mattered. Two preconditions of research having an impact on policy are relevance and timeliness. On the former, although there was an element of luck, two design choices drove success. One was the explicit effort to anticipate issues and launch studies

[29] Many factors explain this shift, but personalities and the change in who controlled Congress likely played a role. During the late l980s when he chaired the relevant subcommittee, Senator Moynihan, a welfare expert and exceptionally nuanced research consumer, consistently sought the latest findings and argued that, given the complexity of the problem, incremental improvements were to be expected (Gueron and Rolston 2013, 199-200). Because of this, I took as a compliment his obviously two-sided description of me as “Our Lady of Modest but Positive Results” (*New York Times,* March 9, 1993).

of enduring policy options. A second was that most of the RCTs did not assess reforms dreamed up by policy wonks. The partnership vision meant that the initiatives tested had bubbled up from governors, their staffs, and community activists — people with finely calibrated judgment on political timing.

On the latter, there is an inherent tension between getting the story out soon and getting it right. Under the state contracts, there was always pressure to produce results quickly, but we were determined not to repeat the negative income tax experiments' experience, where people struggled with limited success to retract findings that had been released prematurely (Coyle and Wildavsky 1986, 179). Yet MDRC faced the reality that it takes time before an adequate number of people are enrolled in a program and can be followed for long enough to judge impacts; it also takes time to obtain, analyze, and accurately report on the data. We sought to address the impatience by dividing up the research: identifying some meaty issues (participation rates, the nature of workfare, implementation challenges) that could be addressed quickly and delaying findings on impacts and cost-effectiveness.

Forceful, nontechnical, and even-handed communication

Finally, people point to the influence of several aspects of MDRC's communication strategy. One was aggressive marketing and outreach to people across the political spectrum. Although this started with lengthy technical reports, it evolved to include pamphlets, press releases, summaries, and more than a hundred presentations — briefings, lectures, frequent testimony — during one year alone. There was also an explicit drive to keep results simple by using easy-to-understand outcome measures and rudimentary and uniform charts and tables that drew, as much as possible, on the transparency of random assignment.

In addition, there was the conscious choice not to take sides and to share positive and negative results.[30] As with many social policy issues, the various factions in the welfare debate differed in their diagnosis of the problem and thus the priority they placed on achieving different goals (for example, reducing dependency or poverty). As a result, good news for some could be judged neutral or bad news by others. MDRC's strategy was not to push people to agree on a goal, but to agree on the facts. Thus, we sought to get reliable estimates of what approaches produced what results and to flag trade-offs, but not to promote or advocate for one policy over another (Gueron and Rolston 2013, 208-211, 443). This style encouraged people with divergent views to see the researchers as neutral parties with no ax to grind.

During the 1980s, the most difficult communication challenge was explaining why, in the face of a competing narrative from prominent governors, high outcomes did not automatically mean success. Staff in states with RCTs begged for cover, as they heard from their own

[30] Although many studies produced positive findings, some were clearly negative. State officials, program administrators, and funders did not welcome reports that progress depended on discarding approaches (particularly their favorite approaches) because they were found not to work. However, though state officials at first may not have grasped that a failed program was not a failed study, we found they did learn and move on from disappointing findings, even to the point of volunteering for subsequent experiments. (See footnote 19.)

governors who — on reading news articles about how other states got tens of thousands of people off of welfare and into jobs — demanded comparably big numbers. How could an RCT suggesting impacts of 5 to 10 percentage points compete? We and the state staff knew from the control groups that most of the people counted in the other states' statistics would have gotten off of welfare anyway, but could they sell that politically? The war of claims played out in the press, but by the late 1980s, after MDRC's relentless outreach effort, key reporters and staff in Congress and Congressional agencies came to recognize that the big numbers could as easily reflect a strong economy as the particulars of welfare reform. However, this was not an argument that was permanently won, and governors continued to duel using competing measures to claim success (Gueron and Rolston 2013, 128-131, 195; Gueron 2005).

This type of active, ongoing communication — to state and federal officials, public interest groups, practitioners, policy analysts, academics, and the press — takes time and money. Throughout these years, MDRC was fortunate to obtain foundation funding to support staff (including communications professionals) in this role. This effort was not viewed as a sideshow, but integral to the organization's two-part mission: to learn what works to improve the well-being of low-income people and to communicate what was learned in ways that would enhance the effectiveness of social policies and programs.

The effect of these four factors was that, despite the highly politicized debate, random assignment was generally accepted as unbiased, impartial, and scientific, rather than as another form of pressure group noise. Further, the findings were not seriously contested and became almost common knowledge. Finally, this result led some people to conclude that the widespread press coverage had an effect on Congress and in states and that the studies contributed to the consensus that made reform possible. As an example, Jo Anne Barnhart, Associate Commissioner/Assistant Secretary of HHS during the Reagan and first Bush administrations, stated:

> The debate over how to reform welfare could aptly be described as contentious, emotional, and partisan. When President Reagan brought his ideas about Community Work Experience [workfare] to Washington, a stark line was drawn in the sand. . . . Without the incremental insights provided by the random assignment experiments, it is difficult to imagine the two conflicting sides coming together. . . . [F]act-based information gleaned from the research provided a "neutral" common language for the divided political rhetoric. Thus, although [the 1996 bill] did not exactly mirror the research findings, it would never have been possible without them. . . . The shift in thinking with respect to welfare reform was the reward [for] the research effort (Gueron and Rolston 2013, 298).[31]

Lessons and Challenges

[31] Reflecting on the 1988 debate, Henry Aaron (1990, 278) offers a contrary view: "The lesson of this experience seems to be that social science can facilitate policy when it finds that measures congenial to the values of elected officials are at least modestly beneficial."

In the field of welfare policy, a long fight showed that random assignment could be used to assess major policy options and that the distinctive quality of the evidence was recognized and valued. This experience provides lessons for others seeking similar rigor.

1. A confluence of supportive factors

In the critical years before 1996, six factors sustained welfare experiments: (1) public hostility to AFDC combined with state/federal cost-sharing to create strong political and financial incentives for governors to innovate and achieve success; (2) the discovery that RCTs were not overly burdensome and could be used to determine the effectiveness of state reforms, plus a growing consensus that alternative methods would fall short; (3) momentum from sufficiently positive findings (success fed success); (4) the active dissemination of results; (5) sustained research funding from Congress, the AFDC formula, and the Ford Foundation; and (6) zealots in the federal government and research firms who stayed involved for decades, consciously built a constituency for experiments, and used the waiver approval process to encourage and ultimately require random assignment.

Researchers in other fields will neither have the same advantages nor have to fight the same battles. The transformation in academic support for experiments is unlikely to be fully reversed and, in combination with the track record of successful RCTs, has contributed to a remarkable federal commitment to scientific, evidence-based policy as a route to more effective government (Gueron and Rolston 2013, 461-68; Haskins and Baron 2011; Haskins and Margolis, 2015). Moreover, as reflected in this volume, hundreds of social experiments are now underway worldwide. The challenge remains to preserve this momentum against future objections and budget cuts and in fields that may be less susceptible to testing.

2. The payoff to building an agenda

The power of the welfare experiments flowed from their logic, relevance, and consistency of findings. In part, this resulted from the independent determination of a small number of people at HHS and MDRC to ensure that successive experiments be accretive rather than a collection of scatter-shot tests (Gueron and Rolston 2013, 431-433). The experiments also responded to the reality of devolution, in which neither the federal government nor any outside actor could impose what would be tested. Welfare reform was too political; the options too controversial. The paradigm of partnership with states, forged out of necessity and that reflected this devolution, had the important benefit of producing results relevant to the diverse and dynamic policy context of state-based welfare programs. Rather than seeking to identify a single, most effective model that no state might have been willing or able to subsequently fund and implement, policymakers pursued evaluations of similar (but not identical) reforms in multiple states as well as a strategically structured agenda that by the end allowed them to see the trade-offs among the major options.

The influence of the experiments also came from the breadth of the research. These experiments were not bare-bones RCTs that spoke only to whether reforms did or did not work. The state and foundation partners would never have gotten involved or stayed the course just for that. They would have found the results insufficiently useful. Although the how and why questions were not answered with the rigor of the yes or no ones (and by how much and for whom), a little insight went a long way toward sustaining momentum and commitment.

Developing this agenda took time. In 1974, it would have been inconceivable to implement RCTs of the scale or complexity of what was done 10 or 15 years later. Researchers did not have the skill or the nerve, nor had they identified the relevant questions. Another reason it took time was that the array of models tested reflected both findings from prior RCTs and the values and beliefs that hardened into policy options after years of debate within states. As a result, constructing the agenda (which eventually encompassed most of the reform proposals advanced during these years) depended on the actual evolution of policy, politics, and evidence.

Over time, there was also a ratcheting up in methodological demands, in terms of the questions asked and the conditions faced. Designs tended to become more ambitious, researchers sometimes had less money and control, and the results became more visible. At each stage, researchers drew lessons on the tools (the art, craft, and risk-taking) they judged key to overcoming the challenges — lessons that were often later revised or reversed.

Implementing this agenda also took long-term funding. High-quality, longitudinal studies (experimental or not) cost money, and the continuity and breadth of the welfare RCTs benefited from there being multiple funders. Most notable, at times when federal enthusiasm waned, support from the Ford Foundation financed the survival of RCTs, the testing of approaches that were of little initial interest to the federal government, and the innovation of the partnership paradigm. Fortunately for those advocating evidence-based policy, there are encouraging signs that public agencies and diverse foundations continue to recognize the vital role they can play in informing policy through supporting rigorous evaluations in the United States and abroad. It remains to be seen whether this role will be sustained and what people and organizations will step up to assure that, as in welfare, the individual studies feed a larger learning agenda.

3. The need for realistic expectations

The welfare experiments tell a surprisingly upbeat story. A range of reforms produced relatively consistent effects: employment rates went up, welfare rolls went down, and there was almost no collateral harm.[32] Some strategies also benefited young children and even substantially reduced poverty. (See Table 3.) Given the skepticism about social programs prevalent in the 1970s — reflected in researchers' fear that the studies would yield null findings

[32] A few studies, however, showed small (sometimes temporary) negative effects on the school performance of adolescents from their mothers' participation in welfare-to-work programs (Gennetian et al. 2002).

(Gueron and Rolston 2013, 45, 205) — and the failure to replicate success in RCTs in other fields, the ability repeatedly to beat the status quo is encouraging.

However, the results also send another message. Average success was generally modest (for example, employment gains of 5 percentage points). Many members of the control groups eventually got jobs or left welfare, either on their own or with the assistance of (or incentives provided by) existing programs and systems. This normal behavior — the counterfactual — set a steep hurdle that reformers had to clear in order to have an impact.

Over the years, defenders of experimental results faced the repeated challenge of setting realistic expectations, especially when politically powerful reformers claimed greater success based on outcomes.[33] But there was one way in which welfare researchers had it easy compared with colleagues in other fields. Reforms that caused people to leave welfare sooner produced real budget savings. Even if controls subsequently caught up, this fading out of impacts did not wipe out past savings. This savings in part explains why almost all states implemented what came to be called “work first” programs.

If RCTs show modest impacts in other fields, will they be viewed as useful building blocks (as is the case for welfare-to-work programs or in medicine) or discarded as signs of failure? Fortunately, the increasing sophistication of public funders (led in the United States by the Office of Management and Budget’s push for high quality evidence of effectiveness) offers reason for optimism. (See Haskins and Margolis 2015.)

4. Maintaining a culture of quality

The welfare experiments were unusual in the extent to which their findings were accepted as objective truth. There were many reasons for this wide acceptance, but two flowed from the shared culture of the relatively small number of people conducting the studies in the early decades. The first was their almost religious devotion to high standards for the myriad aspects that make a quality RCT. The second was their shared vision that the purpose of such studies was to learn *whether* a test treatment worked, not to *prove* that it worked. This eschewing of advocacy research included a commitment to sharing both good and bad news and a view that failure was not learning that a promising program did not work, but of not bothering to learn whether it worked (Gueron, 2008). It is this culture — combined with randomness — that contributed to the view of experiments as the gold standard.

[33] Robert Solow (1980, 16) expressed well the frustration of defending reliably measured impacts against hyped outcomes in his discussion of the Supported Work results: “No one who cares seriously about the employment of disadvantaged groups will sneer at those results. They are not at all trivial. . . . Somehow we have to learn to make a convincing case for policy initiatives based on reasonable estimates of the probability of success and the quantitative meaning of success. If the professional policy community allows itself to promise more than it can deliver, it will end up not delivering what it promises, and eventually the promises will be disbelieved and there will be no delivery at all.”

With social experiments now a growth industry, there is a risk that researchers claim the RCT brand, but do not enforce the multitude of hidden actions vital to the distinctive value of such studies. Just as all that glitters is not gold, the magic does not come from flipping a properly balanced coin. The angel is in the details, and it takes experience to discover and master the details. As policing of RCTs falls to the familiar terrain of peer review, what protects against a debasing of the metal?[34]

5. The advantage of transparent measures and relatively short treatments

People evaluating welfare reforms had several advantages compared with those in some other fields. First, the outcomes that most policymakers cared about — the percent of people working, on welfare, or in poverty and the average dollar earnings or benefits — could be measured in easily understood units (no proxies for what really mattered years later and no hard-to-interpret "effect size") that in most cases could be directly incorporated in a benefit-cost calculation. Second, the treatments were often comparatively simple and short — or usually frontloaded when long or open-ended — so that useful results could be produced with a few years (and sometimes less) of follow-up data. Third, although control group members could and did access competing (and sometimes similar) services provided by other agencies in the community, they were not systematically enrolled in an alternative treatment.

The first advantage had a major impact on communications. At the state level, the studies would likely have had less impact if, at the end, welfare commissioners — who are political appointees — had been told that their programs had an effect size of 0.15 on a measure that was not their ultimate goal (for example, on getting a training credential) and then, in response to the resulting blank stare, been told that this effect was small. My guess is that they would not have acted on the results or volunteered (as some did) to be in another random assignment study. Instead, welfare researchers could make statements, such as "Your program increased earnings by 25 percent and reduced the welfare rolls by 4 percentage points. This cost $800 per person. Over five years, you saved $1.50 for every $1 invested." Since most states wanted to restructure welfare to increase work and save money, this approach was a clear winner. It did not matter that the impacts were called modest or small, the results pointed to a better way to run the system and the response was often direct.

It may be hard to replicate these advantages in other fields, such as education, where the treatments are both more complex and may last many years, the ultimate outcomes are further in the future, the control group members are systematically receiving services, and the goals are more diverse and not convertible to dollar measures. In such cases, studies often rely on intermediate or proximate measures that are an uncertain stand-in for the ultimate goals and are usually calibrated in measures that are not as readily interpretable.

[34] As a warning of the potential seriousness of this risk, Begley and Ioannidis (2015) discuss how the failure to apply well-established guidelines for experimental research may have contributed to the inability to replicate 75 to 90 percent of the preclinical biomedical research published in high-profile journals. In an effort to address this danger, the Institute of Education Sciences created the What Works Clearinghouse to serve as the "central and trusted source of scientific evidence on what works in education" (Gueron and Rolston 2013, 463).

6. The payoff to multiple studies and synthesis

Experience has shown that no single experiment is definitive. Uncertainty shrinks with replication in different contexts and times. The real payoff comes when there are enough high-quality studies to allow for different types of syntheses in order to identify the trade-offs and refine the evidence on what works best for whom under what conditions.

The welfare area was unusual in the extent and nature of experiments and the use of consistent measures. The resulting volume of work and richness of data affected the need and potential for high level syntheses. The result was various kinds of literature reviews, secondary analysis of pooled data, and meta-analyses, including a groundbreaking study by Bloom, Hill, and Riccio (2003, 2005) that applied a multi-level model to pooled data from 69,000 people at 59 offices for which there were identical measures of individual characteristics, management practices, services, economic conditions, and outcomes. (Examples of syntheses include Greenberg and Cebulla 2005, Grogger and Karoly 2005, Gueron and Pauly 1991, Michalopoulos and Schwartz 2001, and Morris et. al 2001.) Among the lessons from this work were that almost all subgroups saw increased earnings from the various welfare reform initiatives, earnings impacts were smaller in places with higher unemployment, and program effectiveness was positively associated with the extent to which staff emphasized rapid job entry and negatively correlated with the extent of participation in basic education (Gueron and Rolston 2013, 348-352).

It will be important to encourage a similar replication of high-quality experiments and uniform data in other fields. (See, for example, Banerjee et al. 2015 and Banerjee, Karlan, and Zinman 2015.)

7. Major challenges remain

The beginning of this chapter posed the fundamental evaluation question: Is it possible to isolate the effect of a social program from the many other factors that influence human behavior? For welfare policy, the answer is clearly yes. Across the country, from small- to full-scale reforms, and under varied conditions, experiments provided convincing answers to the basic question of whether an intervention changed behavior. Moreover, the body of experiments also addressed another question: Is context so important that results cannot be replicated? The answer appears to be no. For reasons that are unclear and in contrast to other areas (Manzi 2012), when the welfare RCTs were repeated (using related, not identical models) in different circumstances, the average results were relatively consistent, providing confidence in the reliability of the findings.

Although the welfare experiments moved the field out of the dark ages of the 1970s, the lack of headway in two key areas suggests some humility. First, despite repeated efforts, the body of work does not adequately explain why programs succeed or fail and thus how to make them more effective. Lurking behind the modest average and broadly consistent impacts is substantial variation. It remains unclear how much of this variation is due to features of people, programs, context, or control services. The uncertainty is not for lack of trying. All the major RCTs used multiple techniques to address this question. Over time, techniques have evolved, including innovative multi-arm tests and the Bloom, Hill, and Riccio study cited above. On-going work promises to move the field further. (For example, see Weiss, Bloom, and Brock 2014 and Bloom and Weiland 2015.)

The second challenge concerns how to make random assignment a more useful management tool. Picking up on what I have stated elsewhere (Gueron and Rolston 2013, 444-446), systematic and repeated RCTs of the type discussed in this chapter provide one view of how to raise performance. It consists of using rigorous and comprehensive evaluations to identify successful approaches, replicating those that work and discarding those that do not, repeatedly modifying and retesting programs, and employing this trial-and-error culling as a means of continuous improvement. Although I endorse this vision, I understand well why critics object to its cost and lag time and also argue that it is too static to serve as a means to foster innovation. There is another approach to using evidence to strengthen social programs: the performance management movement, which sees the real-time tracking of outcome metrics (such as the rate at which people participate or get a job) as a way to achieve multiple goals, including holding managers accountable and inspiring and rewarding progress. Performance management is a bottoms-up approach that sets expectations and leaves managers and staff free to decide how best to use their time and resources to meet or beat the standards.

Ideally, since these two approaches share a common goal of promoting effectiveness by creating a positive feedback loop, they would reinforce each other, with performance metrics serving as a short- or mid-term way to inspire higher outcomes that would, in turn, produce greater impacts and cost-effectiveness (to be periodically confirmed by experiments). But for this result to be true, outcome standards must be a good proxy for impacts. If they are, they will send signals that are likely to make programs more effective; if not, they will increase the risk of unintended, negative effects. Unfortunately, as discussed throughout this chapter, the welfare experiments suggest that outcomes may not be good predictors of impacts. As a result — by making apparent winners out of actual losers — outcomes can potentially send false signals about whom to serve, what managers or practices are most successful, or whether programs are improving over time. (See Heckman et al. 2011.)

This potential for false signals poses a serious dilemma. It cannot mean that outcomes are unimportant, since by definition greater outcomes, if nothing else changes, translate directly into larger impacts. It also cannot mean that workers and managers should not try out and track the results of new ideas unless they are verified by an experiment, since not doing so would deny the obvious value of hands-on experience, high expectations, and incentives. It also cannot mean that setting stretch goals and encouraging people on the ground to figure out ways to achieve them is useless, since that is the way thriving businesses foster innovation and high performance. But it does raise a bright red flag that emphasizing outcomes can prompt people to game the system in a multitude of counterproductive ways. (The press is filled with examples of this response to high-stakes testing in education.)

At present, there is a stalemate, with the two camps existing in parallel. The strengths of one are the weaknesses of the other. Experiments get the right answer about effectiveness but to date have not been useful as a quick turnaround management tool. Outcome standards provide timely and lower-cost data, tap into the "you-get-what-you-measure" mantra, and may stimulate change. But since by definition they measure the wrong thing, the innovation may be implemented in pursuit of a mistaken target.

Over the decades described in this chapter, we have accumulated evidence of this problem but have made only limited progress toward the solution. Although periodic, comprehensive RCTs represent an enormous advance, the challenge remains to more successfully put the tool of social experimentation at the service of managers. One way to accomplish this would be to convince managers to integrate random assignment into their routine testing of small and modest changes in administrative procedures or services, in the process producing treatment and control groups that they or others could follow using existing and low-cost administrative records. This approach resembles the private sector model of rapid and repeated testing, involving hundreds or thousands of RCTs, that Manzi (2012) describes and advocates be applied in the public sector. There is recent interest in this approach, including the creation in 2014 of the first-ever Social and Behavioral Sciences Team in the White House.[35] The concept seems simple, but the tough job remains to convince managers to adopt a culture of evidence-driven innovation and to accept that lotteries are both easy to conduct and a particularly reliable technique to build that evidence. If managers buy into this approach, then rapid-cycle RCTs, with short-term follow-up, could serve as a powerful tool to improve and refine programs, which could then be tested more definitively through comprehensive and longer-term evaluations.

These two challenges are not unique to welfare, pointing to a demanding agenda for future researchers.

References

Aaron, Henry J. 1990. "Review Essay." In "Social Science Research and Policy." *Journal of Human Resources* 25(2): 276–280.

Al-Ubaydli, Omar and John A. List. 2014. "Do Natural Field Experiments Afford Researchers More or Less Control Than Laboratory Experiments? A Simple Model." Working Paper 20877, NBER.

Angrist, Joshua D., and Jorn-Steffen Pischke. 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics." *Journal of Economic Perspectives* 24(2): 3–30.

Banerjee, Abhijit, et al. 2015. "A Multi-faceted Program Causes Lasting Progress for the Very Poor: Evidence from Six Countries." *Science*, 348(6236): pp. 772.

[35] See "A Better Government, One Tweak at a Time," *The New York Times,* September 25, 2015; Social and Behavioral Sciences Team. 2015.

Banerjee, Abhijit, Dean Karlan, and Jonathan Zinman. 2015. "Six Randomized Evaluations of Microcredit: Introduction and Further Steps." *American Economic Journal: Applied Economics* 7 (1): 1-21.

Baron, Jon. 2013. *Statement: House Committee on Ways and Means, Subcommittee on Human Resources Hearing on What Works/Evaluation, July 17, 2013.* Washington, DC: Coalition for Evidence-Based Policy. http://waysandmeans.house.gov/UploadedFiles/Jon_Baron_Testimony_071713.pdf (accessed 11/4/15).

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

Begley, C. Glenn, and John P.A. Ioannidis. 2015. "Reproducibility in Science: Improving the Standard for Basic and Preclinical Research." *Circulation Research*, 116(1), 116-126.

Berlin, Gordon L. 2000. *Encouraging Work and Reducing Poverty: The Impact of Work Incentive Programs.* New York: MDRC.

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

Bloom, Dan, and Charles Michalopoulos. 2001. *How Welfare and Work Policies Affect Employment and Income: A Synthesis of Research.* New York: MDRC.

Bloom, Howard S. (ed.). 2005. *Learning More from Social Experiment: Evolving Analytic Approaches*. New York: Russell Sage Foundation.

Bloom, Howard S. 2008. *"The Core Analytics of Randomized Experiments for Social Research."* In *The SAGE Handbook of Social Research Methods, edited by Pertti Alasuutari, Leonard Bickman, and Julia Brannen*. Thousand Oaks, CA: SAGE Publications.

Bloom, Howard S., Carolyn J. Hill, and James A. Riccio. 2003. "Linking Program Implementation and Effectiveness: Lessons from a Pooled Sample of Welfare-to-Work Experiments." *Journal of Policy Analysis and Management* 22(4): 551-575.

______. 2005. "Modeling Cross-Site Experimental Differences to Find Out Why Program Effectiveness Varies." In *Learning More from Social Experiments: Evolving Analytic Approaches,* edited by Howard S. Bloom. New York: Russell Sage Foundation.

Bloom, Howard S., Charles Michalopoulos, and Carolyn J. Hill. 2005. "Using Experiments to Assess Nonexperimental Comparison-Group Methods for Measuring Program Effects." In *Learning More from Social Experiments: Evolving Analytic Approaches,* edited by Howard S. Bloom. New York: Russell Sage Foundation.

Bloom, Howard S., and Christina Weiland. 2015. *Quantifying Variation in Head Start Effects on Young Children's Cognitive and Socio-Emotional Skills Using Data from the National Head Start Impact Study*. New York: MDRC.

Blum, Barbara B., and Susan Blank. 1990. "Bringing Administrators into the Process." *Public Welfare* 48(4): 4-12*.*

Card, David, Jochen Kluve, and Andrea Weber. 2015. "What Works? A Meta Analysis of Recent Active Labor Market Program Evaluations." Working Paper 21431, National Bureau of Economic Research. Cambridge, MA: National Bureau of Economic Research.

Coyle, Dennis J., and Aaron Wildavsky. 1986. “Social Experimentation in the Face of Formidable Fables.” In *Lessons from the Income Maintenance Experiments: Proceedings of a Conference,* edited by Alicia Munnell. Conference Series 30. Boston, MA: Federal Reserve Bank of Boston.

DeParle, Jason. 2004. *American Dream: Three Women, Ten Kids, and a Nation’s Drive to End Welfare.* New York: Viking Press.

Elmore, Richard F. 1985. “Knowledge Development Under the Youth Employment and Demonstration Projects Act, 1977-81.” In *Youth Employment and Training Programs: The YEDPA Years,* edited by Charles L. Betsey, Robinson G. Hollister, Jr., and Mary R. Papageorgiou. Washington, DC: National Academy Press.

Fraker, Thomas M., and Rebecca A. Maynard. 1987. “The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs.” *Journal of Human Resources* 22(2): 194-227.

Freedman, Stephen, and Jared Smith. 2008a. “Examining the Effectiveness of Different Welfare-to-Work Approaches: Extended Follow-Up of TANF and Employment Outcomes for the National Evaluation of Welfare-to-Work Strategies (NEWWS) Project. Memo 1 -- Long-Term Impacts on Employment and Earnings for the Full Sample and Key Subgroups.” Internal Working Paper. New York: MDRC.

_____. 2008b. “Examining the Effectiveness of Different Welfare-to-Work Approaches: Extended Follow-Up of TANF and Employment Outcomes for the National Evaluation of Welfare-to-Work Strategies (NEWWS) Project. Memo 2 -- Long-Term Impacts on TANF and UI Benefits Receipt for the Full Sample and Key Subgroups.” Internal Working Paper. New York: MDRC.

Freedman, Stephen, et al. 1996. “The GAIN Evaluation: Five-Year Impacts on Employment, Earnings, and AFDC Receipt.” Working paper. New York: MDRC.

Friedlander, Daniel, and Judith M. Gueron. 1992. “Are High-Cost Services More Effective than Low-Cost Services.” In *Evaluating Welfare and Training Programs,* edited by Charles E. Manski and Irwin Garfinkel. Cambridge, MA: Harvard University Press.

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

Gennetian, Lisa A., et al. 2002. *How Welfare and Work Policies for Parents Affect Adolescents: A Synthesis of Research.* New York: MDRC.

Gerber, Alan S., and Donald P. Green. 2012. *Field Experiments.* New York: W.W. Norton and Company.

Glennerster, Rachel, and Kudzai Takavarasha. 2013. *Running Randomized Evaluations: A Practical Guide.* Princeton, NJ: Princeton University Press.

Greenberg, David H., Donna Linksz, and Marvin Mandell. 2003. *Social Experimentation and Public Policymaking.* Washington, DC: Urban Institute Press.

Greenberg, David H., and Mark Shroder. 2004. *The Digest of Social Experiments.* Third edition. Washington, DC: Urban Institute Press.

Greenberg, David H., and Andreas Cebulla. 2005. *Report on a Meta-Analysis of Welfare-to-Work Programs.* Washington, DC: U.S. Department of Health and Human Services.

Greenberg, David H., Victoria Deitch, and Gayle Hamilton. 2010. "A Synthesis of Random Assignment Benefit-Cost Studies of Welfare-to-Work Programs." *Journal of Benefit-Cost Analysis* 1(1): Article 3.

Grogger, Jeffrey, and Lynn A. Karoly. 2005. *Welfare Reform: Effects of a Decade of Change.* Cambridge, MA: Harvard University Press.

Gueron, Judith M. 1990. "Work and Welfare: Lessons on Employment Programs." *Journal of Economic Perspectives* 4(1): 79-98.

_____. 1996. "A Research Context for Welfare Reform." *Journal of Policy Analysis and Management* 15(4): 547-61*.*

_____. 2002. "The Politics of Random Assignment: Implementing Studies and Affecting Policy." In *Evidence Matters: Randomized Trials in Education Research,* edited by Frederick Mosteller and Robert Boruch. Washington, DC: Brookings Institution Press.

_____. 2005. "Throwing Good Money After Bad: A Common Error Misleads Foundations and Policymakers." *Stanford Social Innovation Review*, Fall 2005.

_____. 2008. "Failing Well: Foundations Need to Make More of the Right Kind of Mistakes." *Stanford Social Innovation Review*, Winter 2008.

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

Gueron, Judith M., and Gayle Hamilton. 2002. "The Role of Education and Training in Welfare Reform." *Welfare Reform and Beyond*. Washington, DC: The Brookings Institution.

Gueron, Judith M., and Howard Rolston. 2013. *Fight for Reliable Evidence*. New York: Russell Sage Foundation.

Hamilton, Gayle. 2002. *Moving People from Welfare to Work: Lessons from the National Evaluation of Welfare-to-Work Strategies*. Washington, DC: U.S. Department of Health and Human Services and U.S. Department of Education.

______. 2012. "Improving Employment and Earnings for TANF Recipients." Washington, DC: Urban Institute.

Hamilton, Gayle, et al. 1997. *Evaluating Two Welfare-to-Work Program Approaches: Two-Year Findings on the Labor Force Attachment and Human Capital Development Programs in Three Sites.* Washington, DC: U.S. Department of Health and Human Services.

Hamilton, Gayle, et al. 2001. *How Effective Are Different Welfare-to-Work Approaches? Five-Year Adult and Child Impacts for Eleven Programs.* Washington, DC: U.S. Department of Health and Human Services and U.S. Department of Education.

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

______. 2006. *Work Over Welfare: The Inside Story of the 1996 Welfare Reform Law.* Washington, DC: Brookings Institution Press.

Haskins, Ron, and Jon Baron. 2011. "Building the Connection between Policy and Evidence: The Obama Evidence-Based Initiatives." Paper commissioned by the UK National Endowment for Science, Technology, and the Arts. September. Available at: http://coalition4evidence.org/wordpress/wp-content/uploads/Haskins-Baron-paper-on-fed-evid-based-initiatives-2011.pdf (accessed March 14, 2012).

Haskins, Ron and Greg Margolis. 2015. *Show Me the Evidence: Obama's Fight for Rigor and Results in Social Policy.* Washington, DC: Brookings Institute Press.

Heckman, James J., et al. 2011. *The Performance of Performance Standards.* Kalamazoo, MI: W.E. Upjohn Institute for Employment Research.

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

Hotz, V. Joseph, Guido W. Imbens, and Jacob A. Klerman. 2006. "Evaluating the Differential Effects of Alternative Welfare-to-Work Training Components: A Reanalysis of the California GAIN Program." *Journal of Labor Economics* 24(3): 521-566.

Job Training Longitudinal Survey Research Advisory Panel. 1985. *Recommendations: Report Prepared for the Office of Strategic Planning and Policy Development, Employment and Training Administration.* Washington, DC: U.S. Department of Labor.

LaLonde, Robert. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review* 76(4): 604-620*.*

Manzi, Jim. 2012. *Uncontrolled: The Surprising Payoff of Trial-and-Error for Business, Politics, and Society.* New York: Basic Books.

McLanahan, Sara, and Christopher Jencks. 2015. "Was Moynihan Right? What Happens to Children of Unmarried Mothers." *Education Next* 15(2): 14-20.

MDRC Board of Directors. 1980. *Summary and Findings of the National Supported Work Demonstration.* Cambridge, MA: Ballinger.

Michalopoulos, Charles, and Christine Schwartz. 2001. *What Works Best for Whom? Impacts of 20 Welfare-to-Work Programs by Subgroup.* Washington: U.S. Department of Health and Human Services and the U.S. Department of Education.

Morris, Pamela A., et al. 2001. *How Welfare and Work Policies Affect Children: A Synthesis of Research.* New York: MDRC.

Morris, Pamela A., Lisa A. Gennetian, and Greg J. Duncan. 2005. "Effects of Welfare and Employment Policies on Young Children: New Findings on Policy Experiments Conducted in the Early 1990s." *Social Policy Report* 19(11): 3-18.

Orr, Larry L. 1999. *Social Experiments: Evaluating Public Programs with Experimental Methods.* Thousand Oaks, CA: Sage Publications.

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

Rogers-Dillon, Robin H. 2004. *The Welfare Experiments: Politics and Policy Evaluation.* Stanford, CA: Stanford University Press.

Social and Behavioral Sciences Team. 2015. *Annual Report*. Washington: Executive Office of the President National Science and Technology Council.

Solow, Robert M. 1980. “The Story of a Social Experiment and Some Reflections.” Thirteenth Geary Lecture. Dublin, Ireland: Economic and Social Research Institute.

Szanton, Peter L. 1991. “The Remarkable ‘Quango’: Knowledge, Politics, and Welfare Reform.” *Journal of Policy Analysis and Management* 10(4): 590-602*.*

Weaver, R. Kent. 2000. *Ending Welfare as We Know It.* Washington, DC: Brookings Institution Press.

Weiss, Michael J., Howard S. Bloom, and Thomas Brock. 2014. “A Conceptual Framework for Studying the Sources of Variation.” *Journal of Policy Analysis and Management* 33(3): 778-808.