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

Judith M. Gueron

NBER/MIT
Conference on Economics of Field Experiments
April 10-11, 2015
Cambridge, MA

The Politics and Practice of Social Experiments: Seeds of a Revolution
Judith M. Gueron

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. This 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 studies, and the reanalysis of data from them, as having had an unusual impact on legislation, practice, research methods, and the current enthusiasm for evidence-based policy (Manzi 2012, 181; Baron 2013, 2; Haskins 2006, 11; Haskins and Margolis 2015; de Parle 2004, 111; Angrist and Pischke 2010, 5; Greenberg, Linksz, and Mandell 2003, 238). 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 unfolded.

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 this, I draw, often directly, on *Fighting for Reliable Evidence* (Gueron and Rolston 2013) that centers on MDRC (originally 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 (the sponsor of many of these studies) is obvious; that on a private, nonprofit company makes sense because in the first twenty years that organization conducted many of the major studies 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 as the research director (1974-85) and then president (1986-2004) of MDRC.

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) provide the details behind the headlines, including the critical role

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

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

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 had three mutually reinforcing goals. The first was to obtain reliable evidence of what worked and just as importantly what did not. Over a critical ten years from 1975 to 1985, they became convinced that high-quality RCTs were uniquely able to address such questions 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 emerged 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 programs involving tens of thousands); in the hostility of the context and limitations on centralized control (from testing funded and voluntary services offered by special programs to mandatory obligations in mainstream public agencies); and in complexity (from tests of stand-alone programs to tests of multidimensional system wide reforms using multi-arm experimental designs) – each step of which raised new controversies and objections. This ambition, and a resistance to conducting one-off studies or to evaluating interesting but not central issues, contributed to the long fight to show the feasibility of RCTs under increasingly demanding conditions.

This chapter recounts how, as a result, the challenges, practices, and thus 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 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

[3] I flag this to correct a mistaken view that clout from the federal waiver authority (what came to be called the welfare waiver quid pro quo) explains the flourishing of RCTs. See footnote 13, below.

In the 1970s, knowledge about efforts to move people from welfare to work could accurately be 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 all too often studies of effectiveness ended with experts gathered around a table debating methodology, an outcome that not only was the kiss of death for having an impact on policy makers but also fed the conviction that this research was just another form of advocacy and not "scientific."

The main obstacle to getting persuasive evidence of effectiveness comes from the reality that people on welfare don't stand still waiting for some program to give them a hand up. Many factors influence behavior. Thus, when a woman gets a job, how can one tell if that is because of the help she received, or that the economy improved, or that she got her children into day care, or that she hated the stigma and hassle of public assistance, or some combination of the above plus other reasons? Is it possible for an evaluation to answer that 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 who get a job or a diploma or leave welfare) may accurately tell you their status but will not tell you the change in status that was caused by the program, 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 to their own actions before they enrolled, or to that of people who resembled them on measured characteristics but did not volunteer, or were not selected or served, or lived in a different but similar community. The key 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 that were to occur, the context and/or motivation of people in the two groups 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"), this gives the right answer. On transparency, RCTs allow researchers to estimate impacts using arithmetic. Basically, all one has to do is calculate the average future behavior of people in the two groups and subtract. There are no fancy statistics, no mumbo jumbo of arcane expertise, and scant potential for researcher bias. Anyone and everyone could – and did – understand this simple process.

But the question remained: was it feasible? 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 every day conditions. By the early 2000s, it was clear it was, plus uniquely credible. It was also increasingly clear that alternatives would not get the right answer. How this 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 their 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 included documenting the extent to which the test treatments were implemented (including their operational achievements) and seeking to determine (using varied methods) why they did or did not achieve their goals and what changes would make them more effective.

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 front line staff) with: intake via a lottery; defining and sustaining a distinct treatment; enforcing the research groups (which usually meant not helping the controls) initially and over time; enrolling an appropriate and adequate sample; and cooperating with various research protocols.
2. Obtaining funds for the research and, sometimes, the test program, especially if it was a special demonstration.
3. Gaining the cooperation of the research subjects.
4. Obtaining reliable and comparable data for people in the program and control groups to track outcomes for a long enough time to detect critical effects.
5. Meeting high ethical and legal standards.
6. Assuring that the operating program got a fair test, in particular, that it had 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.

The most fundamental challenge is the first. The researcher needs the agreement of people in the agencies involved. But what is in it for them? Success hinges on an ability to assure them that this alien approach – that for some evokes horrific images of "experimenting" with human beings – is ethical, legal, and actually necessary (that is, that a less intrusive and 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 narrow or secondary questions (Gueron and Rolston 2013, 270-72, 455-68). This was

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

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 their benefit from being in the RCT exceeded these obvious dangers and that, as a result, they wanted you as much as you wanted them. To do this, managers of randomized experiments needed to create a win-win situation. As shown in the rest of this chapter, to do this 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 accept 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 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 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 were tightly controlled and administered by the researchers. The Supported Work challenge would be harder, with much less researcher control, and included convincing 15 mission-driven, community-based non-profit 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 can't be done. Program operators will be implacably opposed to turning people away based on some random process. This will 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,

[5] AFDC, the federal-state cash welfare program created by Franklin D. Roosevelt's New Deal, was eliminated in 1996.

Supported Work was to be a demonstration to 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 this 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, two members of which (Robert Solow and Robert Lampman) backed up by staff at HHS took this in an unanticipated direction by insisting that "testing" meant using random assignment. When asked 35 years later why, Solow attributes his determination to his training – "My first job was as a professor of statistics! I favored it because I wanted to have a defensible response" – and to his and Lampman's conviction that the research design had to be strong enough to detect what they anticipated would be, at best, small and complex effects (quoted in Gueron and Rolston 2013, 32, 483n13).

The result was a hybrid: Supported Work was both a demonstration and an experiment. In its demonstration guise, 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), to allow pooling data from multiple programs, and to reduce the risk of evaluating a poorly implemented start-up period.

Why did ten 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 caseworker discretion, or limiting recruitment so that no one is actually rejected. But staff vastly 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 in Supported Work. But it did not. 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 were given millions of dollars to run a new and distinctive program conditional on their 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 non-financial 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 front line staff stressed the rationale for and morality of random assignment. It was a specially funded demonstration that would provide enriched services that otherwise would not 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. This message was reaffirmed in Supported Work's procedures as the first social experiment to be covered by new federal regulations on the protection of human subjects.[6]

A third factor was the management structure and people involved. Given its complexity, a new organization, MDRC, was created to impose tight central control on the project and explicitly structured and staffed to balance operational and research priorities. MDRC, in turn, selected a team to conduct the impact and benefit-cost analyses that included people at Mathematica Policy Research (MPR) and the University of Wisconsin's Institute for Research on Poverty who had played lead roles in the NIT experiments. This 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.

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 to what was to follow, it was a step out of the laboratory but not a movement into the real world of mainstream public programs. 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.

[6] At intake, through a process of obtaining informed consent, applicants were told about the lottery and the possible risks and informed 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.

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

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

Target group	Program group	Control group	Difference
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. Given the prevailing skepticism, this success was heralded. But this was not the case for the three other groups.

Second, even for the AFDC group, impacts were modest: improvements not solutions. Although Supported Work boosted employment, the big gain over the two years came from the economy and the myriad of other factors that led people (almost all of whom were unemployed at the start of the study) to take a job, as revealed by the employment rate of controls.

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 had welfare as an alternative source of income and would have their benefits 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: the mostly male former addicts, offenders, and 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 controls (without program aid) did so poorly. One implication was clear: traditional outcome-based performance measures (e.g., 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 raised themes that sharpened with time: (1) pay attention to the service differential, that is, don't 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); (2) beware of overreliance on outcome-based performance standards; and (3) look at impacts for key subgroups.

Supported Work also brought good news for 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 substantive pick up. We at MDRC attributed that to several factors: the project's origin (design 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 then-existing systems that would be waiting for the results and primed to act on them. Determined not to repeat that, MDRC took a more inclusive and grassroots approach in subsequent experiments.

Social Experiments Reincarnated as a Partnership: Testing Feasibility Anew

Ronald Reagan's election in 1980 produced a dramatic change in welfare policy, the role of the states, and the nature and origin of research funds. For some time, rising caseloads and shifting attitudes toward women, work, and welfare had fed anger at a system that many felt encouraged dependency, undermined family structure, and unfairly supported people (mainly single mothers) who could work but did not while others struggled at low-wage jobs. As a result, public debate had shifted from whether welfare mothers should work, to who should work and how to make that happen, and from voluntary programs such as Supported Work to mandates and obligations that would require people to work or participate in diverse work-directed activities.

The new administration saw workfare 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. As a result, rather than impose a nationwide vision, 1981 legislation gave the states increased flexibility to undertake their own initiatives. At the same time, the administration, which viewed social

science researchers with suspicion 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. Out of that caldron, it dreamed up a partnership vision that proved to be the major turning point in the design of welfare experiments and within a decade 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 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' famous statement that the states were laboratories for experiments by taking the word experiment literally: that is, by converting into actual RCTs the initiatives that emerged as governors across the country responded enthusiastically to the opportunity to put their stamp on welfare.[7] 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.

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 evaluation. Because the new mandates were intensely controversial, MDRC staff knew they would need the most rigorous evidence to defend any findings. This prompted the choice of 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, ultimately involving 28,500 people. Because of the relatively limited research budget (the Ford Foundation's $3.6 million grant, which MDRC hoped to double, ultimately stretched over more than five years), staff knew they could not track this vast sample using surveys but, instead for the first time in a large-scale RCT, would have to estimate impacts solely from existing

[7] Under AFDC, the open-ended entitlement and federal/state cost-sharing formula gave states 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.

administrative records.[8] This 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-scale bureaucracies, in mandatory programs, with no direct federal funds or role, with no special operating funds, and with no researcher leverage.[9] Further, MDRC would be testing relatively high profile political initiatives that – although still viewed as demonstrations implemented in one or a few locations in the state – were hyped in gubernatorial and even presidential campaigns (one was Governor Bill Clinton's program 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 show 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 two years, MDRC – by making specific design decisions, building relationships that gained trust, and marshalling five arguments to sell the project as a win-win opportunity – gradually overcame these concerns in eight states that met its requirements[10] and as a group were representative of both nationwide responses to the 1981 law and the variety of local conditions.

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 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. They

---

[8] Although this was a decision of necessity, it had the advantage of limiting attrition and recall problems, while raising some coverage issues.

[9] 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 Support 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 sought, in the partnership mode, to be responsive to state policy interests.

[10] 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 somehow provide 50 percent of the funds for the evaluation. This last condition proved by far the toughest.

could grasp how the evidence from control groups in prior RCTs confirmed their doubts. But the challenge remained to explain why one needed an RCT, rather than some less intrusive design to determine success, especially given the limited academic support and often outright opposition. MDRC's response was fourfold: pretend there was a consensus and assume that welfare administrators would not follow the econometric debate; expose the weaknesses of alternative designs; educate them on the outcome/impact distinction and why outcomes would not answer their questions; and offer a study that would accurately measure the accomplishments of their programs, address other questions they cared about (for example, the impact on state budgets and insights on what explained 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 came from a small project (called the Work Incentive [WIN] Laboratories) the organization had managed in the late 1970s 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. However, the 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 this, 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 front line staff would understand the rationale for random assignment; and to produce results that would address pragmatic concerns.

The final selling points were the offer of: a subsidized study that would meet what was then a vague federal requirement for an independent assessment of the waivers to welfare rules that most states needed to implement their initiatives;[11] 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 accrue 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 characterized welfare commissioners as 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) 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, 118, 128-31). In an effort to assist

[11] The subsidy came from the Ford Foundation and, indirectly, federal special demonstration and matching funds. For the critical role of the 50 percent uncapped federal match for state evaluations under AFDC, see Gueron and Rolston 2013.

participating states and debunk these claims, MDRC repeatedly sought to educate the press, advocacy groups, congressional staff, and senior state and federal policy makers 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 results. The result 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 that 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 (eventually more) sought to implement variations on the theme of work requirements. But in the context of welfare RCTs, replication did not mean reproducing an identical, centrally-specified model. Just as welfare benefit levels differed greatly across the country, so did the specific design, targeting, goals, cost, and implementation of the reforms (the message, the participation rates, and the intensity and nature of services). They also differed in context (urban/rural, labor market conditions, and the extent of alternative services). Each state's program having to be a separate RCT created a form of replication that, as discussed below, greatly increased the influence of the findings when 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 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.[12]

[12] As a result, in each site, random assignment was used to create a treatment group subject to the new law and a control group excused from both the newly required services and the threatened financial penalties. 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, 186-8). 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.

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

By 1986, the terrain for welfare experiments had changed. MDRC had shown (as early as 1984) that RCTs testing state initiatives were feasible. A number of senior staff 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 leveraging Ford Foundation grants). At HHS, staff led by Howard Rolston in what was then the Family Support Administration 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. After ups and downs, in 1992 this became the required yardstick to measure the fiscal neutrality of the explosion of waivers that states requested in a push for more, and more radical, reforms.[13]

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

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 statewide programs: Greater Avenues for Independence (GAIN) and Project Independence (PI). Their reasons differed, but neither 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 RCT in San Diego and the usefulness and influence of the findings. The result was that, once they

[13] 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, in order to assure that state reforms did not become an intended or unintended means to drive up the federal cost of AFDC and other entitlements (Gueron and Rolston 2013, 156-59, 217-61).

agreed that GAIN had to be rigorously evaluated, they quickly and across party lines concurred that rigor meant random assignment.

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 work and 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 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 suggested certain approaches were more effective than others	Could this be confirmed in multi-arm RCTs testing varied approaches in the same sites?

Source: Author's compilation

In Florida, after the agency charged by the legislature to determine effectiveness had been attacked for producing conflicting findings from successive studies using non-experimental methods, Don Winstead, the key state official, sought guidance from Robinson Hollister, the chair of a recent National Academy of Sciences panel (see below), 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).[14]

[14] Critically, Winstead had strong support from the Commissioner, Gregory Coler who, notwithstanding negative findings from an 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 RCTs.

Yet, despite strong support at the top and for the first time having random assignment written into the legislation, what followed were legal and ethical objections that went way beyond those raised in the first generation of state studies and, in Florida, produced a firestorm of opposition that almost led the legislature to ban control groups and in the process both jeopardize a major federal research project and potentially poison the well for future studies.

What explains the fierce reaction? The GAIN and 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 and targeted selected groups in a few locations. Now, for the first time, random assignment was proposed to evaluate programs that were intended to be universal (covering all who met the mandatory criteria), full-scale, ongoing, and statewide. Further, the scale was huge. GAIN 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 (with 35,000 ultimately subject to random assignment). This 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? Further, 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, both studies 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 (Gueron and Rolston 2013, 302-04).

People 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. 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 not willing 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." Winstead, when asked why he fought for the study, 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).

A second factor was the slow shift in academic backing for random assignment, reflected in the authoritative 1985 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 an evaluation 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. Further ammunition was provided by influential articles by Robert LaLonde (1986) and Thomas Fraker and Rebecca Maynard (1987) that used data from the Supported Work experiment to show that nonexperimental research designs would yield incorrect conclusions.

A third 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 became particularly important when MDRC and state staff fought to reverse the threat in Florida – using endorsements from these sources and one-on-one meetings with dozens of legislators to sell the merits and ethics of an RCT – out of fear that a successful lawsuit or ban on control group research in that state risked widespread contagion.

The final and most decisive factor in both states was a budget shortfall. Despite the rhetoric of universality, the reality was that there were not adequate funds to serve 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 draw a red line against using RCTs to test entitlements.)

During these same years, HHS launched the most ambitious of the welfare RCTs to evaluate the Job Opportunities and Basic Skills Training (JOBS) program (the major component of the 1988 federal legislation, the Family Support Act) that extended the requirement for participation in work-directed activities to mothers with younger children and emphasized education. The major hypothesis underlying JOBS (as with GAIN) was that remediation of basic skills deficits was central to improving employment outcomes for potential long-term recipients. The GAIN evaluation suggested that this was *not* true: the most successful county was one that emphasized getting a job quickly, but nonetheless provided a mix of activities,

including work-focused short-term education or training (Gueron and Hamilton 2002). This counterintuitive finding along a key liberal-conservative fault line attracted great attention (de Parle 2004, 111). However, since it came from comparing RCT results across California counties (a non-experimental comparison), it cried out for rigorous confirmation.

HHS set out to do this in the JOBS evaluation via head-to-head tests in three sites. Welfare recipients were randomly assigned to a control group or to one of two different approaches: short-term, work-focused activities (called labor force attachment programs) or basic education and related activities (called human-capital-development programs). There were also other multi-arm tests, including using random assignment at two stages in the intake process to measure the separate effect of the program's services and its participation mandate (Gueron and Rolston 2013, 322-31) The evaluation also included an innovative study of the impact of JOBS' work mandate on children in welfare families, an analysis that benefited greatly when identical measures were added to RCTs testing alternative reform strategies (Gueron and Rolston 2013, 331-37, 368-73; Morris et al. 2001; Morris, Gennetian, and Duncan, 2005).

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 complex 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 reluctant partners, the reverse sometimes occurred. Most notably, the Canadian government, the state of Minnesota, and the New Hope program in Milwaukee – all proposing to make work pay more than welfare by supplementing earnings – sought out experimental evaluations as the way to convince a wider audience of the value of their reforms. For them, despite the challenges (particularly in implementing complex multi-arm designs to determine what aspects of their programs drove the impacts), experiments had been transformed from high risk endeavors to a path to recognition. In other states, however, RCTs continued to be accepted grudgingly, as the new price for HHS waivers

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. Fortunately, HHS' commitment to RCTs did not change. After a few years, during which HHS focused on sustaining the most valuable waiver experiments, it shifted gears and took the leadership in launching multi-site experimental projects addressing questions of interest to states in the new TANF environment (Gueron and Rolston 2013, 380-422). 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 (see Gueron and Rolston 2013, 455-71).

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 one that required participation in activities designed to enhance employability – have suggested that the experiments were unusually influential in shaping attitudes, legislation, and practice.[15] None of them claimed that the legislation tracked the RCTs (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 offer four reasons why this group of studies had an outsized influence.

The credibility of random assignment, replication, and relevance

A major rationale for RCTs was the belief that policy makers could distinguish – and might privilege – the uncommon quality of the evidence. For a number of reasons, it seems that this was often the case: the simplicity and transparency of the method; the slowly growing consensus in the research community that alternative designs would fall short; the indication that performance measures such as job placements overstated success (but see below); and the replication of results in diverse conditions, including at full scale. All of these contributed to a bipartisan consensus that the RCTs offered a unusually reliable and objective yardstick.

The reaction to the studies suggested that policy makers also valued external validity, though not in any formal statistical sense. The strategy described earlier -- selecting states judgmentally that were representative along the dimensions politically savvy folks viewed as likely to affect success (e.g., 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.[16]

The findings from comprehensive studies

[15] For example, Ron Haskins, head of the Republican staff on the welfare subcommittee of House Ways and Means during these years, stated: ". . . the experiments . . . had a dramatic effect on the welfare debate. . . . It is the best story I know of how research influenced policy" (quoted in Gueron and Rolston 2013, 297). For different views on how and why these studies did or did not influence policy and practice, see Gueron and Rolston 2013, 190-200, 292-98, 437-44; Baum 1991; Haskins 1991; Szanton 1991; Baron 2013; Weaver 2000, 144; and Greenberg, Linksz, and Mandel 2003.

[16] 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 highlights that the programs were delivered by regular staff in regular offices: "This is no minor matter. In the past, elaborate programs pilot-tested by sophisticated social scientists or a small number of program experts produced worth-while findings. But when the programs were transplanted to real-world social 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" (quoted in Gueron and Rolston 2013, 195).

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.

That the impacts were relatively consistent and in the desired direction (increased work, reduced welfare) was critical.[17] However, the absolute magnitude of impacts also mattered and played out differently in 1988 and 1996. In the early period, the modest gains ("progress" but not "solutions") encouraged expanded funding for welfare-to-work programs; ten years later, this limited success in the face of an increase in the welfare rolls and the more stridently partisan context was taken by some as evidence that a kind of shock therapy was called for.

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 mandates. 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 cost effectiveness measure also provided a useful tool to level the playing field in comparing low- and high-cost reforms.

The timeliness of results

The timing of results also mattered. Although this was in part fortuitous, two design choices drove relevance. One was the explicit effort to anticipate issues and launch studies 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.

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.

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

Although this started with thick technical reports, it evolved to include pamphlets, press releases, summaries, and more than a hundred presentations – briefings, lectures, testimony – during one year alone. There was also an explicit drive to keep results simple: use 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.[18] 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 (e.g., reducing dependency or reducing 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 policy, but to agree on the facts. Thus, it sought to get reliable estimates of what approaches produced what results, to flag any trade-offs, but not to promote one policy over another (Gueron and Rolston 2013, 385, 425). 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 governors who, on reading articles reporting 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 those people would have gotten off of welfare anyway, but could they sell that politically? The war of claims played out in the press, but after a relentless outreach effort, by the late 1980s 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, 128-31, 195; Gueron 2005).

The result of this combination of 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. The findings were not seriously contested and became almost common knowledge. Some concluded 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.[19]

---

[18] Although many studies produced positive findings, some were clearly negative. State officials, program administrators, and funders did not welcome hearing that progress depends on discarding approaches (particularly their favorite approaches) because they were found not to work. And though state officials may not have grasped at first 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.

[19] For example, Jo Anne Barnhart, Associate Commissioner/Assistant Secretary of HHS in 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]

Lessons and Challenges

In welfare, 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 provides lessons for others seeking similar rigor.

1. A confluence of supportive factors

In the critical years before 1996, five 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 to claim success; (2) the discovery that RCTs 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) sustained research funding from Congress, the AFDC formula, and the Ford Foundation; and (5) 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 refight 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 (Haskins and Margolis, 2015). Hundreds of social experiments are now underway worldwide. Nonetheless, it remains to be seen whether these forces and funds will sustain experiments against the next round of 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 showed the independent determination of HHS and MDRC that successive experiments be accretive rather than a collection of scatter-shot tests. It 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

---

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" (quoted in Gueron & Rolston 2013, 298). Henry Aaron offers a useful caution that the findings would not have mattered if policy makers had resisted change: "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" (quoted in Gueron and Rolston 2013, 199).

too controversial. The paradigm of partnership with states, forged out of a necessity that reflected this, 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, the result was both evaluations of similar (but not identical) reforms in multiple states and a strategically structured agenda that by the end allowed policy makers to see the trade-offs among the major options.

The influence of the experiments also came from the breadth of the research. These 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/no ones, a little insight went a long way toward sustaining momentum and commitment.

Building 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 values and beliefs that hardened into policy options after years of debate within states. As a result, building the agenda depended on the actual evolution of policy and politics.

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.

In the future and in other fields, will there be similar drivers and patient funders? Will foundations recognize the vital role they can play in informing policy through supporting rigorous evaluations? High-quality research (experimental or not) costs money, and the continuity and breadth of the welfare research agenda benefited from there being multiple funders. Most notably, at a time when federal enthusiasm waned, long-term 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. Will foundations in other fields take similar risks?

3. The need for realistic expectations

The welfare experiments tell a surprisingly upbeat story. A range of reforms produced relatively consistent effects: work went up, welfare went down, and there was no collateral harm. Some strategies also benefited young children and even substantially reduced poverty. Given the skepticism about social programs prevalent in the 1970s and the failure to replicate success in RCTs in other fields, the ability to repeatedly beat the status quo was encouraging.

However, the results also sent another message. Average success was generally modest (e.g., employment gains of 5 percentage points). Many members of the control groups got jobs or left welfare, either on their own or with the assistance of or with incentives provided by existing programs and systems. This normal behavior—the counterfactual—set a steep hurdle that reformers had to overcome 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. But there was one way in which welfare researchers had it easy compared to colleagues in other fields. Reforms that caused people to leave welfare sooner produced real budget savings. Even if controls eventually caught up, this fade-out of impacts did not wipe out past savings. This in part explains why almost all states implemented what came to be called "work first" programs.

In other fields, if RCTs show modest impacts, will these be viewed as useful building blocks (as is the case for welfare-to-work programs or in medicine) or discarded as signs of failure?

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, 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* the test treatment worked, not to *prove* that it worked. This eschewing of advocacy research included a commitment to sharing good news and bad 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.

With social experiments now a growth industry, there is a risk that researchers claim the brand of an RCT, 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. As policing of RCTs falls to the familiar terrain of peer review, what protects against a debasing of the metal?[20]

5. The advantage of transparent measures and relatively short treatments

[20] As a warning of the potential seriousness of this risk, Begley and Ioannidis (2015) discusses 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" (quoted in Gueron and Rolston 2013, 463).

People evaluating welfare reforms had several advantages compared to those in some other fields. First, the outcomes that most policy makers cared about – the percent of people working, on welfare, or in poverty; the average dollar earnings or benefits – could be measured in easily understood units (no proxies for what really mattered years later, 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 short – or when long or open-ended, were usually front-loaded – so that useful results could be produced from a few years (and sometimes less) of follow up data. Third, although controls 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 program had an effect size of 0.15 on a measure that was not their ultimate goal (e.g., on getting a training credential) and then, in response to the resulting blank stare, been further told that this was a small effect. 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 four 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 was a clear winner. It didn't 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 may last many years, the ultimate outcomes are further in the future, the controls 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.

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 synthesis 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 path-breaking study by Bloom, Hill, Riccio 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 (Grogger and Karoly 2005; Gueron and Pauly 1991; Greenberg and Cebulla 2005; Michalopoulos and Schwartz 2001; Morris et. al 2001; Bloom, Hill, and Riccio 2003, 2005). 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 of the staff's emphasized on rapid job entry and negatively correlated with the extent of participation in basic education (Gueron and Rolston 2013, 348-52).

It will be important to encourage a similar replication of high quality experiments and uniform data in other fields.

7. Major challenges

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, 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 not clear and in contrast to other areas (Manzi 2012), when the welfare RCTs were repeated (not identical models, but related strategies) 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 humility. First, despite repeated efforts, the body of work does not adequately inform 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 is due to features of people, programs, context, or control services. The uncertainty is not for lack of trying. All the major welfare RCTs used multiple techniques to address this question. Over time, techniques have evolved, including innovative multi-arm tests and the groundbreaking 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), systematic and repeated experimentation is one view of how to raise performance: use rigorous evaluations to identify successful approaches; replicate those that work and discard those that do not, keep modifying and retesting programs, and use 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 and ex post to serve as a

means to foster innovation. There is another approach to using evidence to improve 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 improvement. This 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 they share a common goal of promoting effectiveness by creating a positive feedback loop, these two movements would reinforce each other, with performance metrics serving as a short- or intermediate-term way to inspire higher outcomes that would, in turn, result in higher impacts and cost effectiveness (to be periodically confirmed by experiments). But for this 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 effective, or whether programs are improving over time (see Heckman et al. 2011).

This poses a serious dilemma. It cannot mean that outcomes are unimportant, since by definition higher outcomes, if nothing else changes, translate directly into higher 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 this 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 most successful businesses foster innovation and high performance. But it raises 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 and not converging. The strengths of one are the weaknesses of the other. Experiments get the right answer about effectiveness but are not useful as a quick turnaround management tool. Outcome standards provide timely 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 unleashed in pursuit of a mistaken target.

Over the decades described in this chapter, we have accumulated evidence of this problem but have not made progress on the solution. Although periodic, credible evaluations represent an enormous advance, the challenge remains to more successfully put the tool of social experimentation at the service of managers. One route to this would be the integration of random assignment into administrative procedures so that managers (in their routine testing of alternatives or rationing of services) produce counterfactuals that can be used as a lower-cost and more dynamic tool to improve effectiveness.

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

References

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.

Baron, Jon. 2013. *Statement: House Committee on Ways and Means, Subcommittee on Human Resources Hearing on What Works/Evaluation, July 17, 2013,* Washington, D.C.: Coalition for Evidence-Based Policy. [check cite]

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): 603–15.

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

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

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." The Sage Handbook of Social Research Methods….*

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): 551–75.

———. 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, N.Y.: 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, N.Y.: MDRC.

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

DeParle, Jason. 2004. *American Dream: Three Women, Ten Kids, and a Nation's Drive to End Welfare.* New York, N.Y.: 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, D.C.: 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.

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

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

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

Greenberg, David H., and Mark Shroder. 2004. *The Digest of Social Experiments.* 3rd ed. Washington, D.C.: Urban Institute Press.

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

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

Gueron, Judith M. 2005. “Throwing Good Money After Bad: A common error misleads foundations and policymakers,” *Stanford Social Innovation Review*, Fall 2005

Gueron, Judith M. 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, N.Y.: Russell Sage Foundation.

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

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

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

Haskins, Ron. 1991. “Congress Writes a Law: Research and Welfare Reform.” *Journal of Policy Analysis and Management* 10(4): 616–32.

Haskins, Ron. 2006. *Work Over Welfare: The Inside Story of the 1996 Welfare Reform Law*. Washington, D.C.: Brookings Institution Press

Haskins, Ron, and Greg Margolis. 2015. *Show me the Evidence: Obama’s Fight for Rigor and Results in Social Policy*. Washington, D.C.: Brookings Institution Press

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: U.S. Department of Labor.

LaLonde, Robert. 1986. “Evaluating the Econometric Evaluations of Training Programs with Experimental Data.” *American Economic Review* 76(4): 604–20.

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

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 (January).

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

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, Calif.: Sage Publications.

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, D.C.: 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.