Testing Social Policy Using Experiments: Lessons from the United States
Judith M. Gueron, President Emerita, MDRC
Colloque sur les Expérimentations pour les Politiques Publiques de l'Emploi et de la Formation, 23 May 2008, Paris

My remarks and this conference focus on a research method: experimentation. But, in opening, I want to remind us that what we are really talking about is a particular vision of how to make government more rational and effective.

All countries face complex social problems and have options about how to address them. Decision-making is often hard. Many choices create winners and losers. The political process balances different interests to reach decisions. But, are there ways to improve the process? I and other proponents of experiments believe that a key to making smarter choices is getting reliable evidence on whether particular policies and practices do or do not work and at what cost. While having such evidence will not assure that it will be used, it is clear that without it we will often be groping in the dark.

Experimentation makes sense if you view government decisions, at least in part, as emerging from an on-going, dynamic process of testing and refining alternatives, rather than springing full-blown from theory or the discovery of some eternal truth. It also only makes sense if you can actually get reliable evidence. The good news is that over the last 40 years, researchers primarily in the United States have found a way to produce such evidence. Today, I want to briefly share with you the story of how this occurred in one area of social policy, the effort to get people to leave welfare and go to work. As far as I know, welfare reform is unique in having 40 years of uninterrupted, rigorous experimental studies that are widely viewed as having had an important influence on laws and actions..

This is a U.S. story, and even there research is only one of the many factors that influence social policy. I leave it to you to judge whether this vision of improving policy through a process of continuous testing is relevant to France. After all, since its founding, the United States has been called an experiment in democracy. I have never heard people refer to France as an experiment.

Today I will address 5 questions about this history of welfare experiments: (1) How did this happen? (2) What does this teach us about research methods? (3) What does it tell us about what works? (4) Did these studies affect policy? (5) What are some of the big lessons?

Before turning to my subject, I want to share two warnings and a few words on methodology. The first warning is really a reminder. Experiments inform the *process* of decision-making; they do not set the *goals* of policy. From many perspectives, France has social policies that are superior to those in the United States. We may have something to teach you about experiments, but you have much to teach us about social justice. The second warning is that I am not an impartial observer. I spent 30 years helping to build and then leading a non-profit organization, MDRC, that was one of the pioneers in demonstrating the feasibility and power of using experiments to assess social programs.

The methodological issue relates to the crux of the evaluation challenge. I will summarize this briefly to assure that we have a common vocabulary for my later remarks.

Why We Do Experiments

In assessing any social program, you need to address a range of questions.

1. Was the program well implemented?
2. Did it achieve its goals?
3. Are costs reasonable in relation to achievements?
4. Do the answers to these questions vary for different groups of people, policies, and local conditions?
5. What are the lessons for policy and practice?

Today, I am focused mainly on question 2.

Most people find it hard to understand why it is difficult to answer this question. Why can't you simply track people's behavior over time and see if it changes? To understand this, it is essential to recognize the difference between what researchers call "outcomes" and "impacts." A program's outcomes show the status of people at a specific time. Outcomes are such things as how many people (on leaving a program or several years later) get a job or move out of poverty. Impacts are the difference between the outcomes which did occur and the outcomes which would have occurred had the people not been in the program.

The key methodological challenge in evaluating any reform is getting a reliable measure of what people would have done on their own, without the program being tested (what researchers call the counterfactual), in order to determine what a program really accomplished. The problem is, you can never see or directly measure the counterfactual. Politicians and administrators who launch and operate new programs tend to attribute all successful outcomes to their work. For example, they may say with pride that they placed 10,000 people in jobs or moved 5,000 out of poverty. But we know that people don't stand still. Many get or lose jobs all the time. The counterfactual is a moving target. So, is getting 10,000 people jobs a number to be proud of?

How can we determine whether a reform causes a change? How can we avoid repeating the rooster Chantecler's false reasoning that his crowing made the sun rise? It is now accepted by many researchers in the United States that the most reliable method to see whether and how much difference a social program makes is to use a lottery (a process called random assignment) to put individuals or collections of individuals into two or more groups: a program group which is offered or required to be in the special treatment being tested and a control group which is not and provides an estimate of the counterfactual. If the study is well designed and implemented, the difference in their subsequent behavior provides an unbiased estimate of impacts. Such studies are the foundation of evidence-based medicine and increasingly used in evaluating social programs.

Language matters, so I want to make my terminology clear. In casual conversation, people use the term "experiment" to mean trying out something new to see if it "works." When I use the term experiment today, I mean one thing only: a random assignment study. In particular, I do not mean trying out a new idea but measuring impacts with weaker research designs, which many analysts have shown often provide misleading estimates.

I am not going to use my time today to promote the value of experiments, but I do want to encourage you by saying that my personal experience with over 30 major studies involving over 300,000 people in the U.S. and Canada has convinced me that random assignment lives up to its reputation. With experiments you can know something with much greater certainty and, as a result, more confidently separate fact from hype.

The U.S. Context

I will start my story of welfare research with a few words about three aspects of the U.S. context that are relevant to judging the story's applicability to France.

Policy Context: Demand for Change

The first is the strong desire for change. In the U.S., the term "welfare" usually refers to the program of cash assistance for poor, lone parents (primarily mothers) who are not working. The design of this program reflects the shifting priority given to three competing goals: reducing poverty, encouraging parents to support their families, and limiting costs.

In the years I am discussing, many factors made welfare very unpopular, including the rapid increase in program costs and in women's labor force participation. As women – including single mothers with very young children – flooded into jobs, public support evaporated for a welfare program that paid one group of women to stay at home while others were working, often not by choice. The public clearly favored changes that would get people off of welfare and into jobs.

Popular ideas for how to do this included: (1) offering short-term work, training, or other services, (2) requiring people to look for and take jobs, and (3) eliminating financial incentives that discouraged work, an approach that has echoes in the Revenu de Solidarité Active discussion in France.

Decentralized Context: Multiple Sources of Innovation and Money

The second contextual factor is that the U.S. is a highly decentralized country, with multiple sources of innovation and money. Seventy-five years ago, the states were famously characterized as "laboratories" for policy experiments, and this has certainly been the case in welfare, where benefits and rules are partly determined and paid for by the states. As a result, the country is used to variation and ambitious governors, from Ronald Reagan to Bill Clinton, competed to launch successful reforms. Another example is that the entrepreneurs in my story were often not in Washington, but in private foundations and research organizations, a factor critical to sustaining momentum when governments changed or turned against policy research.

Knowledge Context: Demand for Proof

The final contextual factor explains the demand for proof. My story begins in the Dark Ages of the early 1970s, when the U.S. had no answers to the most basic questions about programs to move people from welfare to work.

1. Do they have any effect – positive or negative? If the answer is positive, what is the magnitude? What is the cost?

2. Do impacts vary for different groups of people and types of programs?

3. Is there a tradeoff among program goals, for example between increasing work and reducing poverty?

4. Is the story all about variation, or are impacts replicable in different environments?

5. Can you answer such questions in a way that will be widely believed? In particular, can you do experiments?

6. Are such studies feasible in the real world of large-scale operating programs?

7. Will high quality information make a difference? Can evidence rise above politics and academic squabbles?

The push for answers came from two sources. First, as the number of social programs increased and measured poverty did not decline dramatically, government officials and the public increasingly demanded that social programs prove their effectiveness to earn the right to new or continued funding. Second, advocates for poor people argued that they deserved to know whether the promise that a reform would increase well-being was real or illusory.

Responding to this demand for proof, and as the result of a sustained program of experimental studies, we now have some answers to all seven questions. In describing how this happened and answering the five questions I listed earlier, I will focus on several examples that reflect the evolution of research and policy.

Example #1: The National Supported Work Demonstration

My first example is the 1974 National Supported Work Demonstration, the first random assignment study of a multi-site employment program. Supported Work offered up to 18 months of paid work to four groups of unemployed people – long-term welfare recipients, former drug addicts, people leaving prison, and young school drop outs – with a goal of producing a sustained increase in employment and reduction in behaviors ranging from criminal activities to welfare receipt.

How Did This Happen?

By the mid-1970s, the government had tested many different strategies to help unemployed people find jobs, but the process usually ended in a stalemate. Because of weak research designs – and despite a lot of effort, time, and money – at the end of the study, academics would sit around debating methodology and whether to believe the results. In the U.S., such debates are usually the kiss of death for having an impact on policy.

The entrepreneur behind Supported Work was not someone in government, but a vice president at the Ford Foundation who wanted to find out the potential of a program he had funded in New York City. Since the program was costly, the idea was to learn whether it would work and what it would cost at small scale for tens of millions of dollars before proposing a national program that would cost billions of dollars. To get answers, the foundation recruited federal partners and set up MDRC to make it happen.

In designing the study, we explicitly sought to avoid the legacy of the 1960s, where social policies were often designed on a hunch and discredited on an anecdote, without building a record of reliable evidence of what worked. We argued that, without that record, knowledge could not cumulate and there was a risk that the same strategies would be trotted out every few years, good ideas would be ignored, cynicism would increase, and the country would fail to make progress.

Our solution was to test Supported Work as an experiment. In doing this, we did not naively believe that research would or should drive policy, but we did believe that if you could improve the quality of evidence about effectiveness you would have a chance at achieving several desirable results: improving the lives of low-income people, increasing public support for social programs, and getting a higher return on scarce public investments.

To assure a reliable study, we proposed not only using a lottery but also a large sample, multiple sites, adequate follow up, and high-quality data.

Lessons about Research Methods

What did Supported Work tell us about research methods? Most importantly, it showed the feasibility of using random assignment to evaluate an employment program. With more than 30 years of successful experience, it is easy to get blasé, but at the time this was a revolutionary idea, as it may be now in France. Most people thought that it would be impossible to persuade local program staff to use a lottery to determine who would be served since they would react like a doctor being asked to deny a patient a known benefit and reject the concept as unethical or illegal.

Ultimately, we sold the idea by convincing people that the whole reason for the experiment was that, although Supported Work sounded like a program that could not fail, we did not know whether it would actually help people and, moreover, we had money to enroll only a small number of those likely to volunteer. We argued that, in those conditions, a lottery was the fairest way to allocate these scarce opportunities. We also paid close attention to meeting ethical and legal standards, including getting the informed consent of research subjects and protecting data confidentiality. Finally, the staff promoting the study were not researchers speaking in academic jargon but former program operators who spoke the language of the people we had to convince.

At the national level, we began what has turned out to be a 30-year dialogue about how experiments are not more costly than alternative high quality research methods, but are, instead, more cost effective in that they produce results that can be trusted.

The second methods finding was the feasibility of a survey for tracking thousands of very disadvantaged people for three years and collecting high quality data on sensitive issues including criminal activities.

Third, we were fortunate in that we could measure impacts that were highly transparent and relevant to our audience, for examples, the change in the percent of people working and average earnings.

What Did We Learn About What Works?

The project produced many lessons, but I will mention only a few, because they are counterintuitive. We had expected that Supported Work would have the smallest impact for welfare recipients, since poor women have a harder time finding work than men and have lower work incentives, because they are paid less when they do work and have welfare as alternative source of income. Instead, we found long-term positive impacts for women but not for the three largely-male other groups.

Thus, Supported Work provided the strongest evidence to date that an employment program could have an impact, but also a caution that "good ideas" that seem like obvious winners may not pan out in practice and can actually do harm. We also learned a lesson critical to all of the subsequent studies: People with high outcomes may have low impacts. Thus, the men in the Supported Work program were more likely than the women to find regular jobs, but the men in the control group got jobs just as frequently, which was not the case for the women.

These findings led to what became major themes in these studies: Keep your eye on the control group; programs can have intended consequences; the story is often in the subgroups; it is hard to determine why programs succeed; and without random assignment (that is, just looking at outcomes), we would have reached the wrong conclusions.

Did Supported Work Affect Policy or Practice?

Supported Work showed the multiple ways that research can affect policy. The negative (or, more accurately, null) findings for ex-addicts, ex-offenders, and youth had an immediate effect. The federal government avoided spending huge sums on ineffective employment programs. This is what the planners had hoped, when they espoused testing the program before passing a law.

In contrast, the positive findings for welfare recipients did not lead to an immediate expansion. In my view, this is because this was a stealth project designed in New York and Washington. While this low profile was useful in that it kept the controversial random assignment process below the radar screen, it meant that the state and local political actors, who play a key role in the U.S., did not have any stake in the results. But even this was not the end of the story. Because the findings were from an experiment, they were used in subsequent syntheses of research that did have a major impact on policy. Knowledge did cumulate.

Lessons for Other Fields

Most of the lessons for other fields that I drew at the time concerned the feasibility and value of experiments. It seemed to me then that our success in implementing an experiment was tied to our control of the funding and design of the new program. We were not trying to convince already-funded programs to tack on random assignment (which is always a very tough sell), but instead could insist on it as a condition for getting very substantial operating funds. As a result, we could require some level of standardization and could also assure a large difference in treatment between people who were offered the program and those in the control group, who were not.

A final lesson came from the reanalysis of Supported Work data by a number of researchers that showed that alternative, nonexperimental research designs would have yielded incorrect conclusions.

Example #2: The State Work/Welfare Demonstrations

My second example begins with the election of President Reagan in 1980. This marked a major turning point in this story because it led to dramatic changes in three areas: Policy became more conservative, states were given greater freedom to test ways to require people to work, and the federal government stopped funding most social policy research.

As a result, the prospects for experiments looked bleak. The surprising outcome is that, despite this, experiments not only flowered but within five years shaped a new decentralized paradigm that flourished for the next 15 years and had a greater impact on policy and practice than the experiments carefully nurtured in the more controlled conditions of the 1970s.

How Did This Happen?

When the federal government decided not to study the new state reforms, MDRC got Ford Foundation funds to launch a study designed to answer the three most important open questions: (1) Would states run tough programs? (2) Would these reduce welfare, increase work, or affect poverty? (3) Would such programs cost or save money? Because requiring lone parents to work was highly controversial, we knew we needed the most rigorous evidence to defend any findings and thus proposed using random assignment. Because we anticipated modest impacts and had to assess each state initiative as a separate experiment, we needed very large samples, ultimately involving 35,000 people. As a result, we could not afford surveys, but tracked behavior using existing state administrative records.

In this field, random assignment had never before been done at this scale, in operating welfare offices, independent of the central government in Washington, and without offering states or local programs any special funds.

We sought to recruit states that met a number of conditions. They had to be planning a large new program, have useable data, and collectively be representative of the national response to the new flexibility and of conditions likely to affect program impacts. Further, state governors had to be willing to accept two risks: the potential for backlash and negative publicity from introducing random assignment into the high stress welfare intake process and the possibility that we would produce negative findings for very high profile, political initiatives.

Collectively these features defied the conclusions that I and others had reached from Supported Work about the importance of money and control to the successful implementation of an experiment. So, one could reasonably ask: Why did states participate? The answer is that we very consciously designed and marketed the study as an opportunity for states to answer the questions that they cared about, to receive valuable assistance on program design, and to get visibility by participating in the most important national study.

But candor requires me to say that it was a tough sell. Implementing experiments has almost always been a fight. There was enormous pressure to use weaker, less intrusive research designs. The worst moments I recall are when MDRC staff were called Nazis or accused of using practices akin to those in the most infamous medical travesty in the U.S., the Tuskegee syphilis study.

One reason marketing random assignment was hard is that there was then little support in the universities for this type of work. Quite the contrary, there was widespread and very vocal opposition. Some of it was about statistics, theory, and the legitimacy of different disciplines and views of truth. But part of it was an inter- and intra-disciplinary and organizational competition about who got money and had influence. The rare endorsements from scholars and prestigious panels were absolutely vital to defending experiments as ethical and uniquely reliable.

Lessons about Research Methods

In terms of methods, the studies provided two lessons. They showed that it was feasible (1) to conduct experiments in regular welfare offices and not disrupt normal intake operations and (2) to use existing administrative records to follow people for five or more years and produce reasonably accurate estimates of important impacts.

What Did We Learn About What Works?

In terms of policy, the studies produced numerous important lessons. The state programs requiring people to participate in activities to promote work generally proved successful in increasing work and reducing welfare. Later findings also showed that these mandates did not appear to hurt children in these families. However, average impacts were small to modest, many people remained on welfare and out of work, and there was usually no impact on poverty rates, reflecting in part the way the U.S. welfare system is designed.

Finally, in some states, the programs more than paid for themselves. That is, every dollar invested in operating the program produced several dollars in budget savings. These benefit-cost findings transformed the debate. Suddenly reforms could be described not just as do-good social programs, but as investments with measurable returns.

Did the State Studies Affect Policy or Practice?

Members of Congress and others writing about this period concluded that these studies had an unusual impact on attitudes about welfare and welfare recipients, on the design of state programs, and on federal legislation. They generally attributed this to six factors. The first two − the technical strength of random assignment and the replication of similar findings under different conditions − gave the studies unusual credibility. Basically, no one questioned the findings. The third was timing and relevance: the findings came out in time to affect debates in Congress and the states. Fourth, the programs operated at a scale that was convincing. Fifth, the researchers paid great attention to marketing and communication, and shared both good and bad news. Finally, the political context in the U.S. was less partisan and divisive than it is now. At that time, “modest” impacts were enough to sell Congress on the value of change.

As a result of all of the above, many of the relevant actors by the late 1980s concluded that experiments could be done, would produce results widely viewed as reliable, and would make a difference in policy and practice. This fed an enthusiasm to apply this approach to the remaining big questions.

Example #3: The Next Ten Years

The next 10 years saw a flowering of experiments and the number of researchers involved in these studies. A key reason for this was the U.S. government's insistence on budget neutrality. During these years, state governors proposed increasingly radical ideas, from putting a limit on how long someone could receive welfare, to generously supplementing the earnings of working families, to providing more expensive education and training. The governors wanted flexibility, not research, but for the first time federal officials declared that state reforms could not increase federal budgets. Moreover, they insisted that the yardstick for assessing budget neutrality would be a random assignment study. As a result, agreement to an experiment became the quid pro quo for state flexibility.

I have time to share only three findings from these years. The first is evidence of the risk from using outcomes to judge program success. While outcome standards (measures such as the percent of people placed in jobs or at wages that get them out of poverty) can be a useful tool to motivate managers, they can also prompt programs to make inefficient decisions. High outcomes may reflect not program success but a strong local labor market or the enrollment of more motivated participants. An experiment can distinguish these, but an emphasis on outcomes alone may prompt managers to change whom they enroll rather than to improve what they do.

The second finding was that no single approach did best on all the policy goals: increasing work, reducing poverty, saving money, improving outcomes for children. There were trade-offs. As a result, ones conclusion on "What works best?" will depend on what goals one cares most about. For example, as I said, programs that required people to get jobs quickly increased employment, saved money, and did not harm children, but also did not reduce poverty. In contrast, programs that supplemented the earnings of welfare recipients who took full-time jobs increased work, reduced poverty, and had a positive affect on the school performance of young children in these families, but they cost more.

The third was about methodology. A series of studies used data from the welfare experiments to show the failure of alternative research designs to replicate the results. It was these studies, combined with the evidence of the feasibility and persuasiveness of random assignment, which over these 40 years made reluctant converts of many of us to a strong belief in the unique virtues of experiments.

From Research To Policy: Lessons From the U.S. Experience

At the beginning of my remarks, I described experiments as a means to make government more effective. In the United States, the welfare story is viewed as a model of how research can inform decisions. In conclusion, I would like to offer 12 lessons from this experience for others who seek to use experiments to improve policy.

Lesson 1: Address important issues. The life cycle of a major experiment is usually three or more years. To be successful, the study should address issues that matter – and that will still be of interest when the results come in – and about which there are important unanswered questions.

Lesson 2: Have a reasonable treatment. An experiment should test an approach that is supported by past research and looks feasible operationally and politically, by which I mean that it is likely that the relevant administrative systems will cooperate, that people will participate enough for the program to make a difference, and that the costs will not be so high as to rule out expansion.

Lesson 3: Design a real-world test. The program should be tested fairly (if possible, after the initial start-up period) and, if feasible, in multiple sites, since we have learned that context matters. It would be uniquely convincing to be able to say that similar results emerged in Saint Etienne, Bordeaux, and Paris.

Lesson 4: Address the questions that people care about. Does the program work? For whom? Under what conditions? Why? Can it be replicated? How do benefits compare with costs? Are there trade-offs among goals?

Lesson 5: Fight for random assignment. A high-quality random assignment study is superior in providing a reliable estimate of whether a program works. Yet, if France is anything like the U.S., proponents of experiments will have to overcome resistance from administrators, politicians, and even some other researchers who argue that such studies are some combination of unnecessary, unethical, illegal, burdensome, or unreasonably expensive. The easy response is to accept a weaker design, but the second best study is often not worth the paper it is written on. While it is important to acknowledge those situations when random assignment cannot be used or will not answer the right questions, these are less numerous than opponents will claim.

To implement experiments successfully and get the most out of them, experience suggests being careful to meet ethical and legal standards, remaining sensitive to local concerns, and combining different research methods to examine which features of the program or its implementation account for success or failure.

Lesson 6: No single experiment is definitive. One study cannot address all questions for all time. Certainty cumulates with replication. In social policy as in medicine, the real payoff is when there are enough high-quality, experimental results to allow for various types of synthesis in an effort to identify the trade-offs and find out what works best for whom under what conditions. In the U.S., sustaining a long-term program of experiments required building a community of funders, researchers, advocates, and the media that valued and could distinguish high quality studies.

Lesson 7: Do not define success as working miracles or you are likely to fail. To sustain a multi-year program of experiments, you need some good news. The welfare studies delivered this, in that they found many reforms that produced positive changes. But they also showed that the magnitude of change was modest. For example, employment rates might increase seven or ten percentage points, earnings increase 25 percent, and welfare roles decline five percentage points. These impacts represent clear progress, but not miracle cures. This is less a caution about experiments than evidence of how hard it is to change behavior, of the importance of the economy, and that the policies tested were often not dramatically different from the services available to people in the control group.

For this audience, I would ask: In France, are reforms likely to produce bigger impacts? If not, would findings of modest success be viewed as building blocks to progress (as in medicine or sometimes in welfare in the United States) or as evidence of failure?

Lesson 8: Simplify. One of the beauties of an experiment is that anyone can understand what you did. There are no fancy statistics. In our experience, if an advanced degree is needed to understand the lessons, they are not likely to reach policymakers. We built on this simplicity in multiple ways. We presented the results in a standardized format; we used the same simple outcome measures in all the studies; and we usually put equations in appendices. We also avoided overly complex research designs, although complexity did increase over time. But we did not avoid telling people that the findings were complex, or involved tradeoffs, or needed to be understood in the context of other research.

Lesson 9: Actively communicate the results. Politicians and funders are impatient consumers. The welfare projects were structured so that some findings came out in a year or two and were aggressively shared with the multiple interested audiences, including the news media. At the same time, there was a conscious effort to resist pressure to produce results so early that subsequent findings risked reversing the conclusions.

Lesson 10: Do not confuse dissemination with advocacy. The key to long-term successful communication is trust. If you overstate your findings or distort them to fit an agenda, people will know it and will ultimately reject what you have to say. The researchers' role is to learn whether something works, not to prove that it works.

Lesson 11: Be honest about failures. Let's face it, public officials and program operators share the human fondness for good news. They don't welcome hearing that progress depends on identifying and discarding approaches that do not work, and that their program is one of those. To their credit, however, we have found many people able to learn and move on from disappointing findings.

Lesson 12: Get partners and buy-in from the beginning. In the U.S., I would add a final lesson that, in conceptualizing and launching a project, it is important to involve the major actors and interest groups from the beginning so that they understand and have a stake in the research and are less likely to attack the methods or the findings. This is also likely to improve the study.

These 12 lessons emerge from a U.S. context of skepticism about whether social programs work. I leave to you, the experts on France, to judge whether they apply here. Beyond that, it is always wise to retain some humility. While it is not necessarily pleasant, researchers should remember that their work is only one ingredient in the policy process and that, when the stakes are high enough, at least in the U.S., politics trumps research. Our job is to bring truth to power. Experiments are vital to doing that. But power resides elsewhere.

Thirty years ago, European policymakers and scholars who visited me at MDRC typically had a two-part reaction when they heard about social policy experiments. The first response was awe and admiration: “You are really testing social programs as doctors test new drugs? That’s amazing.” Their second response was relief: “Thank God we don’t have to do that. When we want to adopt a new policy, we can just pass a law.” This conference suggests that the atmosphere here may be changing. I hope it proves to be for the better.

Thank you.