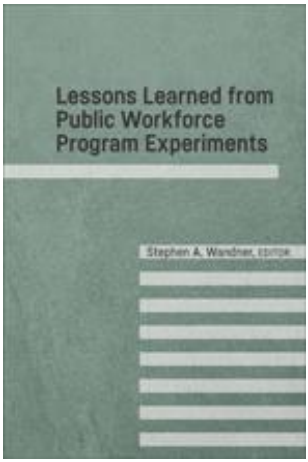
W.E. UPJOHN INSTITUTE
FOR EMPLOYMENT RESEARCH

Upjohn Institute Press

How On-the-Ground Realities Shape the Design, Implementation, and Results of Experimental Studies



Irma Perez-Johnson
American Institutes for Research

Annalisa Mastri
Mathematica Policy Research

Samia Amin
American Institutes for Research

Chapter 2 (pp. 13-40) in:
Lessons Learned from Public Workforce Program Experiments
Stephen A. Wandner, editor.
Kalamazoo, MI: W.E. Upjohn Institute for Employment Research, 2017.
DOI: https://doi.org/10.17848/9780880996310.ch2

Chapter 2

How On-the-Ground Realities Shape the Design, Implementation, and Results of Experimental Studies

Irma Perez-Johnson
Annalisa Mastri
Samia Amin
Mathematica Policy Research

Planning and implementing a large-scale experimental evaluation of a social program is not unlike planning and embarking on a major road trip. Before leaving, we carefully plot the route, identify likely stopping points, book needed accommodations, and even check for road construction and other potential obstacles along the way. Similarly, a significant amount of effort is invested up front in the design and planning for the launch of an experimental study. One specifies the intervention's theory of change or logic model, identifies the outcomes of interest, determines necessary sample sizes, specifies random-assignment procedures, identifies the data sources and analytic methods that will be used to evaluate results, recruits study sites, and generally tries to plan for all the needed details and anticipate as many roadblocks as possible.

Despite all this planning, the one certainty in both major road trips and experimental studies is that one will encounter unanticipated challenges and will have to adapt quickly. Evaluators need to balance the ideal with the practical while maintaining analytical rigor. For instance, when conducting a study of a program being implemented in multiple sites, often the ideal would be for all study sites to offer identical services to clients, delivered by staff with similar

backgrounds and training, and with the same level of resources. But this is rarely feasible.

This chapter draws on our experiences designing and implementing three experimental studies of social programs to discuss how "on the ground" realities can shape the design, implementation, and results of such studies. We first provide some background on each study, then discuss considerations for designing, executing, and interpreting the results of such studies. We conclude with a summary of lessons that can inform similar efforts moving forward.

THREE ILLUSTRATIVE EXPERIMENTAL STUDIES

We draw on our experiences from three large-scale experimental studies sponsored by the U.S. Department of Labor (USDOL): 1) the Individual Training Account Experiment (ITA Experiment), 2) the Workforce Investment Act Adult and Dislocated Worker Programs Gold Standard Evaluation (WIA Evaluation), and 3) the Self-Employment Training Demonstration (SET Demonstration). The ITA Experiment and the WIA Evaluation were conducted within the context of ongoing programs under the Workforce Investment Act of 1998 (WIA). The SET Demonstration project was designed to show proof-of-concept—that is, to test whether a new program could be implemented with fidelity to the model and achieve the desired effects. Although all three studies took place in workforce-related settings with individual job seekers, many of the lessons learned from these experiments can be applied more broadly to experimental studies of human services programs.

The Individual Training Account Experiment

The ITA Experiment examined the effectiveness of three alternative models of delivering WIA-funded training vouchers, known as ITAs. Although WIA directed states to restrict available training pro-

grams for local high-demand occupations, it also gave states considerable flexibility in structuring their ITAs. For example, states could vary the amount of money they offered to trainees, the counseling supports offered, and the amount of counseling they required trainees to complete before getting access to an ITA (Perez-Johnson et al. 2000).

The ITA Experiment sought to determine the optimal combinations of training dollars and counseling supports by testing the following three approaches:

1) **Guided choice**. This option featured fixed-amount and moderately sized ITAs ($3,500 on average) and some mandatory counseling activities to guide trainees' program choices. Guided choice was designed to resemble the widespread practice used in the early 2000s, when the study was launched (Trutko and Barnow 1999).

2) **Structured choice**. In this option, trainees could receive customized ITAs with a higher cap (around $7,500) but were required to complete more intensive counseling, and caseworkers could veto training choices on which they expected to have low labor-market returns.

3) **Maximum choice**. In this option, customers received the same ITA amount as under guided choice. Counseling to discuss training options, although available, was not required. Trainees had to request counseling if they wanted it.

The experiment took place in eight Local Workforce Investment Areas (LWIAs) across the United States. A total of 7,920 participants were randomly assigned to one of the three ITA approaches (Perez-Johnson et al. 2004). Staff in each LWIA worked with customers in all three study groups to avoid staff-specific effects on participants' outcomes, and all participants in each study site had the same menu of training programs to choose from. The experiment compared participants' training and earnings outcomes for up to seven years after entry into the study (Perez-Johnson, Moore, and Santillano 2011).

The WIA Adult and Dislocated Worker Programs Gold Standard Evaluation

The WIA Evaluation (Mastri et al. 2015) aimed to estimate the relative effectiveness of three tiers of services offered through the WIA Adult and Dislocated Worker Programs:

1) **Core services**. Available to all customers, core services typically include self-service activities such as accessing job listings and local labor market information in a resource room or on the Internet.

2) **Core and intensive services**. Customers who are unable to get a job that would lead to self-sufficiency using core services alone may access intensive services. These services can include customers working with a counselor to develop an employment plan and obtain in-depth assessments of their skills, interests, and abilities.

3) **The full WIA offering (core + intensive + training)**. Customers who need a skills upgrade to obtain or retain employment can request ITAs to fund training from approved providers.

The WIA Evaluation was designed to produce nationally representative impacts. Therefore, the study first randomly selected LWIAs nationwide and then convinced them to participate in the study. In these LWIAs, almost all customers who requested and were eligible for WIA intensive services or training and who consented to participate in the study were randomly assigned to one of the three groups described above. Study intake occurred between November 2011 and February 2013, with intake durations varying between 2 and 16 months across the participating LWIAs. Across the 28 LWIAs that participated in the study, 35,665 customers were randomly assigned to one of the three groups. The evaluation examined the service receipt and labor market outcomes of study participants measured at 15 months and 30 months after their enrollment in the study.

The Self-Employment Training Demonstration

The ongoing SET Demonstration is testing strategies to support dislocated workers who want to start their own businesses (Amin et al. 2016). Unemployed and underemployed workers who propose establishing businesses in their fields of expertise are eligible for the program. Eligible applicants are randomly assigned either to a treatment group that gets access to SET services, in addition to whatever other services are available in the local area, or to a control group, which cannot access the SET program but can seek out the other services available in the area. The treatment group participants can receive up to 12 months of counseling, training, and technical assistance on business development from experienced providers, as well as up to $1,000 in seed capital microgrants to help them establish their businesses. As noted, the SET Demonstration was designed to illustrate proof-of-concept of a new program rather than evaluate existing services. Thus, the SET program itself had to first be developed alongside the evaluation, then sites selected to implement it. The program is being tested in four metropolitan areas across the United States (Chicago; Cleveland; Los Angeles; and Portland, Oregon). Enrollment occurred between July 2013 and February 2016, yielding a sample of 1,981 study participants. The evaluation will use survey data to measure SET's impact on receipt of self-employment assistance services, self-employment experiences, employment and earnings (both from self-employment and from wage/salary jobs), and job satisfaction. It is also drawing on qualitative data from site visits, phone interviews, and a management information system to examine program implementation.

HOW ON-THE-GROUND REALITIES SHAPE THE DESIGN, IMPLEMENTATION, AND RESULTS OF EXPERIMENTAL STUDIES

With the key features of these studies in mind, we now discuss how real-world factors frequently affect the design, implementation, and interpretation of results of experimental studies. The design phase encompasses everything from specifying the research questions of interest to conducting power calculations and identifying and recruiting sites that are suitable for the evaluation. The implementation phase includes a period of training for participating sites in study procedures, the study intake period, and the period of data collection and analysis. The results phase focuses on interpreting the study's results in light of the design and implementation experiences. Finally, we discuss special considerations for demonstration programs such as SET.

Study Design

The first phase in which on-the-ground realities begin to shape the experimental study is during the study's design phase. Key considerations include the following:

Selecting policy-relevant treatment contrasts

The study will have the most potential to detect program impacts if it can compare program receipt with the absence of similar services (i.e., a no-treatment control condition) or compare treatments that differ notably along key dimensions (i.e., contrasting important alternative approaches). A no-treatment counterfactual may be infeasible for an experimental study in the case of mandated or entitlement programs that do not allow denial of services to eligible individuals. In these cases, an experimental design could involve randomly assigning additional program components on top of entitled services, but it could not deny entitled services to form a control group. For instance,

WIA mandates universal access to core services. Therefore, the WIA Evaluation had to be designed so that all study participants could receive core services, and the study randomly assigned those who could receive intensive counseling and training services on top of core services. This made it impossible for the study to determine the overall effectiveness of WIA relative to no access to any WIA service.

Depending on the study's goals, comparing alternative approaches may be the preferred design even if a no-treatment contrast is feasible. For instance, in the ITA Experiment, the goal was to determine the optimal approach to structure ITAs, rather than to assess the net impacts of ITA training. In that case, it did not make sense to have a control group with no access to ITAs. For the WIA Experiment, the treatment-control contrasts would have been greater if study participants didn't have access to Wagner-Peyser Act employment services, which in some LWIAs are very similar to intensive services. But restricting study participants' access to Wagner-Peyser services would have represented a counterfactual that does not exist in reality, making the study's results less policy relevant. Similarly, in the SET study, the objective was to learn whether having the opportunity to participate in SET would result in better outcomes than having access to the usual infrastructure for business development support. In this scenario, it did not make sense, nor was it feasible, to ask partner providers or others to refuse their usual services to members of the SET control group.

Researchers and the policymakers interested in the studies' results must recognize in advance the potential limitations of the study findings. For example, in the ITA Experiment, a finding of no differences between the study groups would not imply that training was ineffective, just that the various voucher approaches did not change customer outcomes. Similarly, the SET study can show whether *enhanced* self-employment services matter, but not whether self-employment services in general are effective. Null findings from alternative treatment contrasts must be interpreted carefully, and a no-service control group can typically answer a broader set of policy-relevant questions.

Matching data collection plans with desired indicators

Ideally, the outcomes of interest to the study align well with the anticipated effects of the intervention being tested and are already captured in administrative data sources. However, these conditions frequently do not hold. For instance, in the SET Demonstration, the intervention was expected to affect self-employment activities and success rates of new businesses started by study participants. But data on these key outcomes of interest are not currently contained in administrative databases. Self-employment activities are not reportable to the state unemployment insurance reporting system, which is the typical source of administrative data on people's earnings. Thus, the study team decided early on that a survey of study participants would be necessary to collect key outcomes.

Collecting information on the treatment group's service receipt is important for interpreting the study's results. However, programs' management information systems do not always collect service receipt data at the level of detail ideal for the study. Moreover, it is rarely the case that these systems collect data on services that treatment-group members receive outside the program or on any services that control-group members receive. In addition, program staff often differ in the extent to which they update administrative systems with such information. For the SET Demonstration, participants' service receipt and achievement of program milestones were critical components of illustrating proof-of-concept. Therefore, the study team offered service providers financial incentives to record information on participants' progress through SET in a study-designed tracking system.

Finally, collecting information on control group members' service receipt is often important in understanding the counterfactual condition and interpreting the study's results. This is especially true in service-rich environments, where control-group members might be receiving services that are very similar to those of treatment-group members. This was the case in the SET study, and surveys are usually the best source of these data.

To be sure, there are some instances where the context in which the program operates facilitates use of existing data sources. For instance, a program implemented in a workforce system setting among a segment of job-seeking customers might be able to use administrative data the workforce system is already collecting on all customers to examine the employment outcomes of participants. Care must be taken in the design phase, however, to assess not only the feasibility of using existing administrative data sources but also the quality of these data, and to develop a backup plan if appropriate.

Short-listing good candidates for the study

Not all sites are good candidates for inclusion in an experimental study. Some are in the midst of making big changes to their programming or organizational structures. Others may be reluctant to participate in an experimental study, or are already participating in a different study. And some sites will simply lack the client flow necessary for participating in a study.

For the WIA Evaluation, we began with a list of the full set of LWIAs nationwide. Given the evaluation's target sample sizes and intake period, we excluded LWIAs serving fewer than 100 customers per year. Not having adequate client flow can be even more of an issue for experimental studies of grant programs that might only award enough funds to serve a relatively small number of customers in each site, meaning that many sites would need to participate in the study in order to detect meaningful impacts. In some cases, there simply might not be enough total participants over the study's time frame to support an experimental study.

Recruiting enough suitable sites to participate in the study

Having an adequate number of sites participating in the study is critical for being able to detect meaningful program impacts. However, recruiting sites is rarely a straightforward process, even when there are a number of potentially suitable sites for the evaluation. For

the ITA Experiment, USDOL issued a request for proposals to interested LWIAs or consortia of agencies, and only six expressed interest in implementing the experiment, despite generous financial incentives to do so. For the SET Demonstration, the study team preselected six metropolitan areas in six states that met the necessary conditions for the study and conducted targeted outreach to them. LWIAs in one metropolitan area were keen to participate but did not have sufficient provider capacity to implement the model as planned. LWIAs in another metropolitan area did not want to participate because of concerns about staff burden.

The WIA Evaluation faced an additional challenge. To be nationally representative, the study needed to first randomly select LWIAs nationwide and then convince them to participate in the study. We aimed to include 30 LWIAs in the study and successfully recruited 26 (87 percent) of these. In addition, we recruited two others that were randomly selected to replace two of the four local areas that declined to participate, for a total of 28 local areas participating in the study.

For all three studies, our recruitment efforts succeeded because we did the following:

We involved all key stakeholders in recruitment presentations and meetings. Having all interested parties at the table is important for achieving buy-in. For the WIA Evaluation, we first had to identify who the relevant stakeholders were in each LWIA. In some, the LWIA could not agree to participate without support from the state, so state representatives were involved. Some potential sites asked us to present in front of their local boards. Others asked us to meet with line staff or staff from partner organizations or referral sources. For SET, we conducted a series of recruitment e-mails and calls—first with regional and state officials from both the UI office and the workforce development departments. We followed up these initial contacts with meetings with the local workforce agencies. We wrapped up with in-person visits to each of the promising sites.

We demonstrated that we had the strong commitment and support of USDOL. For the WIA evaluation, federal USDOL staff participated in recruiting trips, the assistant secretary of labor sent letters to LWIAs and also made phone calls to some of the more reluctant sites, and USDOL staff held a special session at a conference attended by LWIA representatives to discuss the importance of the study. For SET, senior USDOL staff wrote e-mails to invite participation and participated in site recruitment calls.

We offered compensation. The WIA, ITA, and SET studies all provided sites with payments to offset the staff time associated with implementing study procedures and cooperating with evaluation activities. These payments could be used, for example, to hire an additional staff person to assist with data entry or documentation needed for the study (e.g., assembling and securely storing study forms), or to hire an additional staff person to handle the added caseload.

We offered concessions to make the experiment more attractive to sites. The design phase is a good time to start thinking through parts of the experiment that might be contentious from the standpoint of the sites. Even when sites are mandated to participate in an experiment (for instance, as a condition of receiving grant funding), researchers want to ensure their buy-in for the study so that they maximize the study's potential success. Common sticking points include the following three:

1) **Denying services to customers in the control group(s).** Program staff are motivated by a desire to help their customers and often find the idea of denying services unpalatable. They also might be reluctant to deny services because it could result in unused capacity at their programs. A potential solution to both problems is to reduce the ratio of participants that are denied services. Although having equally sized treatment and control groups is optimal from a statistical power standpoint, sites may be more comfort-

able with randomization if a smaller proportion of participants are turned away. For the WIA Evaluation, we determined the total sample size needs with an average random-assignment rate of around 6 percent each to the core and core-and-intensive groups. The other 88 percent of customers were randomly assigned to the group that could receive full WIA services as usual. However, these study group assignment rates were possible because of the large study sample size.

2) **Implementing burdensome study procedures.** If sites perceive that they will have to engage in burdensome documentation or other study procedures, they will be less likely to cooperate. During the design phase, steps can be taken to reduce the study's burden on participating sites. For instance, for the SET Demonstration, we developed an online orientation video and online participant application to save American Job Center staff from conducting these in person. The study team also planned to review applications, make eligibility determinations, conduct random assignment, and refer individuals accepted into SET to the local microenterprise providers who provided services. For the WIA Evaluation, staff at American Job Centers had to conduct the random assignment, so we developed an easy-to-use online random-assignment system that would only require staff to enter minimal information as data.

3) **Concern about the effects on the program's performance.** Program administrators may worry about how the study will affect the extent to which they meet or exceed legislatively mandated performance targets. Depending on the study design, this could occur because the programs will be serving fewer total customers (since some will be assigned to the control group) or because some of the particularly promising customers—who would have contributed favorably to the program's performance reporting—will be

assigned to the control group and hence not be included in the reporting. In the case of demonstration programs, desired outcomes for the new project might not align with the existing performance measures. Efforts can be made to see whether program offices can relax or adapt performance requirements for the duration of the study.

Implementing the Study

Often, we can anticipate during the design phase many of the challenges of implementing a social experiment in the context that we are studying, and we begin adapting the study design at that point. But ultimately the success of the evaluation hinges on how well the evaluation team adapts and responds to challenges encountered during the evaluation's implementation. Key challenges we frequently have faced and to which we have had to adapt include the following:

Achieving buy-in of program staff

Even after sites have agreed to participate in a study—a decision that is often made at the administrative or higher level—frontline staff working directly with participants might still not understand the value of an experimental study. It is particularly of concern when the control group will not have access to study services and cannot find alternative services in the community. Sometimes program staff have such dedication to serving their clients that they are resistant to participating in the study or following study procedures.

Explaining the importance of the study in easy-to-understand terms is critical to achieving staff buy-in. For the WIA Evaluation, we developed a presentation and accompanying one-page fact sheet aimed at line staff and supervisors. We delivered the presentation during site recruitment visits and distributed the fact sheets throughout the course of the evaluation. Both described the study's importance, goals, and basic outlines of the design in layman's terms. In particular, they centered on why random assignment is the strongest research

design, why it is ethical, and how the study's results would be used to make future funding decisions, allowing program staff to continue or even expand their good work.

Building capacity of program staff to implement study procedures correctly

Program staff are typically not accustomed to explaining a research study to potential participants, obtaining their consent to participate, entering information into a tracking or random-assignment system, conducting random assignment, and informing study participants of their study-group assignments. Yet successfully implementing each of these steps is critical to the success of the experimental study.

To address these challenges for the WIA Evaluation, the study team thoroughly investigated the program service delivery structure, customer flow, and staffing, and developed study procedures customized for each site in order to be as seamless as possible with existing service delivery. The study team developed customized study procedure manuals documenting in detail each step that program staff had to take for the study. The manuals included explanations of the study forms, scripts for explaining the study to customers and collecting their consent, and detailed information on how to use the random-assignment system. Information was presented in multiple ways—for instance, as scripts and as talking points that staff could adapt to their own style. We conducted a day-long training at each site for staff to go over the study procedures in detail and to allow staff the opportunity to practice role playing. Early training sessions revealed that, in our zealousness to provide sites with lots of details and many options for conveying information to study participants, some staff felt overwhelmed by the volume of information provided. So we developed a reference guide that boiled down the study procedures manual into a 10-page document that staff could more easily reference on a daily basis.

The SET Demonstration took a different approach. The study team limited LWIA responsibilities to referring potential clients to

the SET website so they didn't have to deal with complex study procedures. As described above, the study team handled orientations, applications, random assignment, and client referral to services. Nonetheless, since only one out of our four sites was familiar with self-employment services, we still had to build LWIA capacity for promoting this new kind of program. To do so, we provided detailed procedure manuals, in-person training, scripts for describing the SET program, and attractive brochures and fliers to help promote SET.

We learned over time, however, that just simplifying the procedures and initial training was not sufficient. One unintended side effect of not involving LWIA staff in orientations and intake for SET was that they were less familiar with the program and less invested in its success. Capacity building therefore became an ongoing effort. We discuss below how we created feedback loops to address this issue.

Reserving resources to maintain buy-in and capacity for the evaluation

Commitment, energy, and attention to the evaluation may wane over the course of implementation. Initially, sites may be excited about what they can learn through participation, or—in the case of demonstration projects—about the prospect of offering new services to their customers. However, it takes enduring commitment on the part of program administrators and staff to follow through on that initial excitement. This is especially true in times when site partners are under strain—when resources run low, staff turnover is high, or staff face many competing responsibilities. In those circumstances, it can be asking a lot for sites and their staff to maintain their commitment and attention to a temporary initiative. It can be especially challenging in the case of demonstration projects in which a whole new program must be tested (see Box 2.1).

Supporting program staff to correctly implement study procedures cannot end with training. Study teams must devote resources to providing ample support throughout the study enrollment period and extra technical assistance when support and effort appear to be wan-

Box 2.1 Special Considerations for Demonstration Projects: Ensuring Program Fidelity

The goal of a demonstration project is twofold: first, to provide "proof of concept" (i.e., a demonstration that a program can be successfully implemented while being faithful to the model) and, second, to evaluate the impacts of the program. Project staff must be prepared for the reality that *it can take time to implement the demonstration program with fidelity*. Program staff often struggle to deliver a new set of services as planned, especially when they are quite different from what they are used to delivering. This can result in delays in beginning to offer services in the first place, and in weak program implementation.

In some cases, monitoring to identify implementation difficulties and providing ongoing support to the sites can rectify these issues. For instance, early in the SET Demonstration, the study team noticed through visits to the SET service providers that case management services—a critical element of the program—were not being delivered as frequently or thoroughly as intended. The study team provided technical assistance on the case management model to 5 of the 11 service providers over three to eight months. The study team also initiated monthly phone calls to monitor implementation fidelity and provide an opportunity for SET service providers to ask questions of the evaluation team.

In other cases, the demonstration (or some aspect of it) proves difficult to implement because it is too much of a departure from standard practice. For instance, the Structured Choice approach in the ITA Experiment was not fully implemented because counselors felt uncomfortable vetoing customers' training choices. The study authors concluded that a substantial cultural shift would need to take place for program staff to successfully implement the Structured Choice approach (as originally envisioned) to administering ITAs. Notably, despite the more limited implementation of the Structured Choice approach in the ITA Experiment, it proved the most effective of the three ITA approaches tested (Perez-Johnson, Moore, and Santillano 2011).

ing. Training must be followed by monitoring phone calls with program staff and, if possible, site visits to understand the implementation challenges and determine how best to address them. These issues could range from the relatively insignificant, such as assembling forms in the wrong order, to the significant, such as not randomly assigning everyone who is eligible for the study. Designating a person from the evaluation team to serve as a liaison with each site can be an effective way to monitor implementation and handle questions or concerns from program staff; all three of our studies provided this. In addition, both the WIA Evaluation and the SET Demonstration operated hotlines for staff to call with questions.

For the SET Demonstration, the study team realized that, because SET was a new and temporary program (available for less than three years) and not directly provided by LWIA staff, workforce staff were promoting it less and less over time than they had at the beginning of operations. Moreover, budget pressures in the participating LWIAs were making it difficult for overburdened workforce and UI staff to focus on SET. The team introduced feedback loops to get the buy-in of the referral sources in the LWIAs. We sent monthly e-mail updates to all LWIAs on the progress of their recruitment efforts relative to other sites. This encouraged sites that were recruiting well to keep up the good work and motivated some of our lagging sites to become more competitive. For sites where recruitment was lagging, we conducted in-person visits to retrain and motivate staff and followed that up with biweekly calls. Sharing client success stories and testimonials proved to be a particularly effective strategy but was only feasible once the program had been in operation for a while. It helped to generate excitement about SET and made staff more comfortable in referring clients to the new program.

Increasing assistance when staff burden exceeds expectations

Except in rare cases, some interruptions to the ideal flow of services that staff provide and participants receive should be expected. Staff members, in addition to their regular duties, must perform study

procedures and, in many cases, some data entry in support of the study. During the study implementation phase, they might find that implementing study procedures is taking longer than the evaluation team had anticipated, imposing unexpected burdens on program staff and causing errors in following study procedures.

One way to address any unanticipated burden is to offer additional compensation or other resources to local partners. For the SET Demonstration, for example, we provided additional funds to support special outreach activities such as mailings in sites where meeting recruiting targets required additional effort. For all of our sites, we provided additional supplies of publicity materials (fliers, brochures) whenever they were needed because these were expensive for our partner sites to produce. We also provided ongoing support by designating evaluation liaisons to each study site to help troubleshoot emerging challenges.

Developing new tools to minimize the burden on staff can also help. During the WIA Evaluation, some sites noted that introducing and explaining the evaluation was taking them substantially longer than anticipated. As a result, the evaluation team developed a video—much like the one used in the SET Demonstration—that staff could play for customers either at their desks or in a group orientation setting. This freed up staff to work on other tasks while the video played.

Another example was adding study group assignment to existing program management information systems. Because program staff conducted random assignment for the study, they had to enter some information about customers into an online random-assignment system we had designed for the study, and customers' study group assignments were recorded there. Although we had designed the system so that minimal data entry was required, some staff complained that they had to look up every customer with whom they met in the online system to see whether they were already enrolled in the evaluation and, if so, what services they were allowed to receive. They noted that it would be easier to look up this information in their existing management information systems, which they would be accessing

anyway when working with customers. As a result, the evaluation team worked with data systems personnel at the state level to add fields with which to document study groups within the existing statewide management information system.

Documenting variations in service delivery across participating sites

Differences in staff backgrounds and training, program context, participant characteristics, and the way service providers are accustomed to delivering services mean that services that are nominally the same may not actually be delivered in exactly the same way across sites. For instance, WIA gives local areas considerable discretion in service delivery. We found that during the WIA Evaluation a core "job search workshop" varied in length from a couple of hours at one site to three days at another. It also was categorized as an intensive service at some sites but a core service at others.

In the ITA Experiment, the goal was for participating sites to implement the same three ITA approaches. However, ITA caps had varied across sites before the study, and sites needed to set the caps high enough that they would spend their entire training budget (or lose it the next year). As a result, the caps for each treatment arm necessarily varied across sites. Other variations included which occupations were considered high wage and high demand, whether assessments were required and used as a counseling tool, and supervisor involvement in the approval of customer training selections under the Structured Choice approach.

The SET Demonstration was designed to provide a common service flow across microenterprise service delivery providers, including individualized service planning, monthly check-ins, quarterly reassessments and service plan updates, and the $1,000 seed capital microgrant available to participants who met required milestones. Within these parameters, however, sites varied in how they structured their check-ins, the degree to which they relied on workshops and group classes, and the range of technical assistance and additional

services they offered to SET participants. The infrastructure for self-employment support—e.g., the number or reach of individual providers and the overall culture of entrepreneurship—also varied across sites.

Documenting these variations is important for interpreting the findings of the impact analysis and providing lessons learned for program improvement. Large experimental evaluations often have an implementation study tied to them—particularly in the case of demonstration projects—in which qualitative researchers systematically collect information on many aspects of program organization, operations, and staffing, among other topics. This can be a rich source of information on variation in program delivery across participating study sites. Lower-cost methods such as phone calls with sites and online staff surveys can also be good sources of this information.

Addressing changes in service delivery in response to the study

Studies of ongoing programs would ideally examine the effectiveness of services as they are typically delivered. However, sometimes a service offering or its delivery changes in unexpected ways as a result of the study. In the ITA Experiment, some private training vendors appeared to change the content and price of their offerings in response to the study, bundling additional certificates together and charging a higher price because the ITA cap was higher for some customers as a result of the experiment. During the WIA Evaluation, staff in some local areas reported that referral sources such as local community colleges were "drying up" because of a misperception that the local area was no longer funding ITAs. And in evaluations where study enrollment is lower than anticipated, assigning a fraction of study participants to a control group may leave some slots unfilled. As a result, staff may find that they have more resources to serve any given customer, thereby allowing them to deliver more intensive services than they would in the absence of the study.

The study team must pay attention to these issues for the duration of the study enrollment and follow-up period and address any sub-

stantial changes in service delivery to the extent possible. Failure to do so could have significant implications, particularly if the program is so changed that it no longer provides an accurate picture of how the program will operate once the study is over. In other words, the evaluation is of a program that does not exist. In the case of the ITA Experiment, little could be done to change service provider prices or restrict participants from asking for ITA funds up to the cap available to them. For the WIA Evaluation, the study team worked with managers and administrators at the LWIA to coordinate outreach to their referral sources to explain the study and emphasize that training funds were still available.

Monitoring sample sizes and adjusting procedures accordingly

Recruitment often lags behind what was anticipated based on previous history or projected customer flow. If sample sizes substantially lag behind projections, the study will be less able to detect meaningful program impacts. Based on historical data on the number of customers served in participating LWIAs, the WIA Evaluation expected to enroll about 85,000 customers in the study. In actuality, only about 36,000 were enrolled. In the SET Demonstration, enrollment lagged substantially below targets, partly because the eligibility requirements were fairly narrow and partly because the high unemployment levels that prompted the demonstration in the first place had abated by the time the program began. In the ITA Experiment, the opposite happened—the economic downturn that occurred around the time of the study, which was unanticipated, increased the overall flow of customers and trainees through the participating LWIAs, resulting in much larger sample sizes over the study's two-year implementation period.

It is imperative that researchers monitor sample buildup and work with sites to understand and rectify, to the extent possible, problems with recruiting enough participants. The SET Demonstration revised outreach materials to make them simpler and more accessible to potential participants. They worked on achieving buy-in from LWIA staff so that they would spread the word and promote the dem-

onstration program. They also tailored outreach tactics for each site, including boosting advertising efforts in some sites and even hiring a marketing firm for outreach in one site.

The WIA Evaluation encouraged program staff to tap their referral sources, but the main approach to combating lower-than-expected sample sizes was to adjust the rates at which customers in lagging sites were randomly assigned to the core or core-and-intensive groups. (As mentioned earlier, originally only 12 percent of all WIA customers were supposed to be referred to these two groups.) In short, achieving the target number of customers in these groups was crucial to maintaining the study's power. Since the total number of customers was lower than expected in some sites, we had to increase the proportion of customers assigned to these groups. In addition, the study enrollment period for some of the participating local areas was extended beyond the originally planned 12 months to allow for additional sample buildup. In the end, the study met its enrollment targets for customers assigned to the core and core-and-intensive groups.

Interpreting the Study's Results

The flexibility and adaptations the study team makes in response to the realities of designing and executing an experimental study in the context of social programs have implications for the interpretation of the study's results. Key challenges include the following:

The counterfactual is weaker than anticipated in some sites

The services available to control group members in the broader community often vary across sites, sometimes substantially. If control group members in sites with many similar alternative services make use of those services, the differences between the treatment and control groups are narrowed. This makes it more difficult for the study to detect impacts of the program. For example, in some LWIAs participating in the WIA Evaluation, there was no other public source of training funds available, whereas in others, alternative sources of

training funds were readily available. This meant that the treatment-control contrast was weaker in the latter sites, again making it harder to detect program impacts.

If the suitability assessment conducted during the design phase identifies sites where a lot of alternative services are available, the study team can consider excluding such service-rich sites. Sometimes, however, the extent of alternatives available is not known until after the study is launched and data are collected. In those instances, the evaluation's results must be interpreted through the lens of what was actually being tested on the ground, and not in theory. Doing this requires investigating and documenting the services available to the control group and, if possible, capturing control-group service receipt. When statistical power allows, impacts can be investigated by site and compared among those with strong treatment-control differentials and those with weaker ones.

The sample size is lower than anticipated

Strategies to increase recruitment and enrollment in programs are not always effective, and sample sizes can fall short of targets. There are limited options available in this scenario. If the analysis was planned to be done by site, the data analysis could instead pool the sites to boost statistical power. However, the conceptual model of the program and the research questions of interest would have to suggest that pooling could be appropriate. For instance, it might not make sense to pool sites with completely different service delivery strategies and target populations, even if they are funded by the same grant.

In these scenarios, conducting post hoc power analyses is useful for determining the magnitude of effects the study can detect, given its realized sample sizes. This can help policymakers and others better interpret the study's findings. For instance, a large but statistically insignificant point estimate could indicate a lack of statistical power, rather than the lack of a true impact of the program on the outcome of interest.

Sample attrition is high

Sometimes it can be difficult to locate study participants for follow-up data collection efforts. This is especially true when the program under study targets a hard-to-reach population, such as homeless people or formerly incarcerated adults. When experiments have high overall study attrition, or large differences in attrition rates between the treatment and control groups, the amount of bias in the impact estimates rises, making us less confident in the study's results. The What Works Clearinghouse, a systematic evidence review project funded by the U.S. Department of Education, developed a bias model that specifies the combinations of overall and differential attrition that are acceptable in experimental studies. Studies that exceed the specified thresholds are considered to have a high likelihood of biased impact estimates.

In cases of high overall or differential sample attrition, a carefully controlled analysis is one approach to reducing bias in the estimated impacts. Ideally, the authors would include controls for the demographic characteristics of the study sample and preprogram measures of the outcomes that are of interest to the study. In a study examining the impact of a job training program on earnings, this preprogram measure could include the earnings history of participants leading up to the point of random assignment.[1] Another approach is to explicitly demonstrate that the study participants included in the analysis sample (i.e., those for whom follow-up data were available) were similar in preprogram demographics and outcomes at the time of random assignment. This can be done by performing statistical tests of the baseline differences between the study groups. However, ultimately, high attrition of study participants cannot be "controlled away," and many evidence reviews would downgrade such studies.

LESSONS LEARNED

In this paper, we've sought to draw on experiences in designing and conducting three large-scale experimental studies to discuss how on-the-ground realities influence the design, implementation, and results of these studies. In many cases, a flexible and adaptable approach to the evaluation can mitigate the issues encountered. We hope that the lessons we have learned from these and other evaluations can inform future efforts. Specifically, we offer the following advice:

- When selecting sites for the evaluation, think carefully about the objectives of the study and the characteristics of the sites that could potentially participate. Are they strong or weak implementers? Do they likely have sufficient sample size? What is the availability of similar services in the community? Balance the needs for representativeness and for evaluating the program as closely as possible to how it would operate in the absence of the study against the need for feasibility in successfully implementing the study at the site.
- When recruiting sites to participate in an experiment, prepare easy-to-understand materials about the goals and benefits of the study and its ethics. Have in mind concessions that could be offered to the sites to minimize the impact of study participation on their service delivery. Make sure all stakeholders are at the table during the recruiting process.
- If possible, involve federal sponsors of the study during site recruitment and throughout the course of the evaluation to demonstrate a commitment to and support of the study.
- For demonstrations, recruit sites with a strong commitment to and interest in the concept being tested, as well as interest in learning from the study's results. Build feedback loops for staff and program administrators to learn how the demonstration program operates in practice, what services are provided to referred clients, and what benefits participants derive from

the opportunity to participate. This information is critical to sustain enthusiasm for the program and a commitment to offering the program to suitable candidates.

- If study resources permit, compensate sites for their time and effort spent implementing the study—this helps to achieve buy-in, facilitates site recruitment, and lessens the burden on site staff.
- Develop both detailed manuals and easy-to-use resources to support implementation; quick reference guides are key. For demonstrations, be clear on the elements that must be preserved and those that can be adapted; aim for flexibility wherever possible.
- Reserve resources to provide lots of training and ongoing support for the study sites. Designate a site liaison to facilitate communication about the evaluation, monitor site progress early to correct any mistakes, and monitor sites on an ongoing basis to ensure they maintain their focus on and fidelity to evaluation procedures. Adapt the frequency and intensity of monitoring as needed. Implementation issues can evolve and change over time, especially in the context of a multiyear program. For instance, staff turnover, the business cycle, and spikes or severe dips in application rates or program referrals can all affect study implementation at various points in time.
- Take proactive steps to minimize the burden on local staff. For example, automate procedures to the extent possible and, if feasible, embed data collection into existing management information systems. Provide ready-made resources for staff, such as promotional brochures and posters, run help lines to handle questions about the program, and designate a single point of contact from the study team to handle questions or concerns from local staff. Share information about burden-reducing and other facilitating strategies or resources that participating sites develop on their own.

- Document variation in program implementation and services available to the control group to help interpret results.
- If, despite the adaptations made along the way, issues remain with the implementation of the experiment, conduct supplementary analyses when possible and discuss the results. For instance, conduct post hoc power analyses if sample sizes are low, change the analysis approach if implementation was not strong in some sites or the treatment-control contrast was weak, and include controls or demonstrate baseline equivalence if attrition was high. Although these methods might not answer the original research questions of interest with the level of rigor originally intended, they can still provide meaningful answers to important questions about the effectiveness of programs under study.

Note

The projects discussed in this chapter were funded, either wholly or in part, with federal funds from the U.S. Department of Labor, Employment and Training Administration. The contents of this chapter do not necessarily reflect the views or policies of USDOL, nor does mention of trade names, commercial products, or organizations imply endorsement of same by the U.S. government.

1. Some systematic evidence reviews, including the Clearinghouse for Labor Evaluation and Research, require that earnings and employment history be measured for more than one year before random assignment to guard against the Ashenfelter dip.

References

Amin, Samia, Heinrich Hock, Irma Perez-Johnson, Shawn Marsh, Mary Anne Anderson, and Rob Fairlie. 2016. *Evaluation of the Self-Employment Training Demonstration: Design Report*. Princeton, NJ: Mathematica Policy Research.

Mastri, Annalisa, Sheena McConnell, Linda Rosenberg, Peter Schochet, Dana Rotz, Andrew Clarkwest, Ken Fortson, AnnaMaria McCutcheon, Katie Bodenlos, Jessica Ziegler, and Paul Burkander. 2015. *Evaluating National Ongoing Programs: Implementing the WIA Adult and Dislocated Worker Programs Gold Standard Evaluation*. Submitted to the U.S. Department of Labor, Employment and Training Administration. Washington, DC: Mathematica Policy Research.

Perez-Johnson, Irma, Paul Decker, Sheena McConnell, Robert Olsen, Jacquelyn Anderson, Ronald D'Amico, and Jeffrey Salzman. 2000. *The Individual Training Account Demonstration: Design Report*. Washington, DC: Mathematica Policy Research.

Perez-Johnson, Irma, Sheena McConnell, Paul T. Decker, Jeanne Bellotti, Jeffrey Salzman, and Jessica Pearlman. 2004. *The Effects of Customer Choice: First Findings from the Individual Training Account Experiment.* Princeton, NJ: Mathematica Policy Research.

Perez-Johnson, Irma, Quinn Moore, and Robert Santillano. 2011. *Improving the Effectiveness of Individual Training Accounts: Long-Term Findings from an Experimental Evaluation of Three Service Delivery Models.* Princeton, NJ: Mathematica Policy Research.

Trutko, John W., and Burt S. Barnow. 1999. "Experiences with Job Training Vouchers under the Job Training Partnership Act and Implications for Individual Training Accounts under the Workforce Investment Act." Unpublished paper. Washington, DC: U.S. Department of Labor, Employment and Training Administration.

Lessons Learned from Public Workforce Program Experiments

Stephen A. Wandner
Editor

2017

WE *focus* series

W.E. Upjohn Institute for Employment Research
Kalamazoo, Michigan

Library of Congress Cataloging-in-Publication Data

Names: Wandner, Stephen A., editor.
Title: Lessons learned from public workforce program experiments / Stephen A. Wandner, editor.
Description: Kalamazoo, Michigan : W.E. Upjohn Institute for Employment Research, 2017. | Series: WE focus series | Includes index.
Identifiers: LCCN 2017044711 | ISBN 9780880996303 (pbk. : alk. paper) | ISBN 0880996307 (pbk. : alk. paper)
Subjects: LCSH: Public services employment—United States. | Manpower policy—United States. | Unemployment—United States.
Classification: LCC HD5713.6.U54 L47 2017 | DDC 331.12/0420973—dc23 LC record available at https://lccn.loc.gov/2017044711

W.E. Upjohn Institute for Employment Research
300 S. Westnedge Avenue
Kalamazoo, Michigan 49007-4686

Cover design by Carol A.S. Derks.
Index prepared by Diane Worden.
Printed in the United States of America.
Printed on recycled paper.