Measuring Success in Pay for Success

Randomized Controlled Trials as the Starting Point

Justin Milner and Kelly Walsh

August 2016

As the number of pay for success (PFS) projects continues to grow in the United States and around the world, a central question remains: What is the best way to measure success in PFS endeavors? In the context of PFS (box 1), success refers to a social program's ability to improve outcomes for a specific group of people. This is not a "business-as-usual" approach for government. For most social programs, examination of the links between the actual work of the program—such as homelessness prevention or after-school supports—and the impact it has upon participants simply does not take place. Yet funds invested in rigorous program evaluations are minimal (Haskins and Baron 2011), meaning government often has no way of knowing whether a program is effective.

For PFS projects, however, evaluation comes to the forefront. With investment returns riding on the demonstration of program results, determining the best approach to measuring success is a critical challenge for many PFS stakeholders ranging from mayor's offices, to community service providers, to government agency officials, to philanthropic and private investors. Regardless of their individual incentives for participating in a PFS project, all of these actors are dependent on an accurate evaluation of program outcomes. If a program succeeds, governments repay investors, investors may earn a return on investment, service providers benefit from an objective validation of their model, and program participants' lives improve. The measurement of success has definitive consequences for all PFS participants.

So how do you measure success in PFS? A good starting point for discussions on evaluating PFS projects should be randomized controlled trials (RCTs). According to government and scientific

authorities, RCTs are the strongest way to evaluate program effectiveness.[1] When conducted well, they have a significant strength: "[RCTs] are often the best, and sometimes the only, way to draw conclusions about an intervention's causal impact," (Theodos et al. 2014). In other words, RCTs enable one to assess whether the program itself, as opposed to other factors, actually caused the observed outcomes. Though RCTs may not be the appropriate method for all research projects, their strengths are particularly well suited to the requirements of a PFS project.

This brief provides a basic overview of RCTs to help PFS stakeholders engaged in project development who may lack formal evaluation expertise. The brief will describe the PFS model and the design of an RCT, explore the advantages of RCTs and why they are particularly useful for PFS projects, and address a few key questions on RCTs that project partners may raise.

BOX 1

What Is PFS?

Pay for success, or PFS, is an innovative financing mechanism that shifts risk for a new but evidence-based social program from a traditional funder (usually a government) to a third-party investor (usually a private organization or nonprofit). At the heart of all PFS projects is a test of whether a social program can improve outcomes for a specific group of people. If the program works (as measured by a rigorous evaluation), the project is a success. Investors get their money back (with a potential positive return), the government realizes potential future cost savings, families and society benefit from better outcomes, and social service providers strengthen the case for funding their model. For more information on PFS, visit https://pfssupport.urban.org/

What Are RCTs?

An RCT, also referred to as an experimental study, is a type of evaluation generally considered the most reliable way to determine a program's impact (Tatian 2016).[2] Discerning the true impact of a program allows researchers and PFS stakeholders to make definitive conclusions about whether a program or some other set of factors helped improve the lives of a specific group of people.

RCTs are experiments designed to answer a single question: Will people (or classrooms, communities, or other units of interest) who participate in a program[3] experience greater beneficial outcomes than people who do not? In an RCT, evaluators randomly assign participants to two conditions: the treatment group (who receive program services) and the control group (who receive business-as-usual services). The evaluation is controlled to make sure that participants have an equal probability of being assigned to the treatment group. The control group provides a benchmark, or counterfactual, to understand the net impact of the program; in other words, the control group reflects what would have happened to the treatment group in the absence of the program (Gueron 2002). Over the course of the evaluation, researchers track performance for both groups. At the end, they compare

the outcomes for the treatment group to those of the control group. The difference is the impact of the program.

Why Should PFS Projects Build RCTs into Their Design?

In the context of PFS projects, there are several reasons why stakeholders should strongly consider using an RCT to measure success. The results and repayments in early PFS projects have been closely scrutinized, and current projects continue to attract significant public attention as examples of innovative government work. This creates a strong incentive for PFS project partners to use the best possible methodology for measuring success, reducing points of possible critique. RCTs enable us to easily present results, provide a fair way to distribute access to a service when resources are limited, balance a focus on precision during project formation with a simple analysis on the back end, and generally offer the best way to build strong evidence on a program's effectiveness.

Further, although other methods can over- or underestimate program impact, RCTs generally provide the most accurate reflection of the program's effect by comparing performance to a randomized control group (box 2). The method is not without limitations, particularly with regard to how much one can generalize its findings,[4] but the strengths of RCTs are particularly well suited to the requirements of a PFS project.

RCTs enable us to easily present results, provide a fair way to distribute access to a service when resources are limited, balance a focus on precision during project formation with a simple analysis on the back end, and generally offer the best way to build strong evidence on a program's effectiveness.

BOX 2

RCTs Reduce Bias in Estimating Program Impact

In 1996, researchers published an evaluation of Even Start, a federal program designed to improve literacy and other outcomes. The method chosen was a pre-post study that analyzed outcomes experienced by participants before the program and outcomes experienced by them after, attributing any changes between the two to the impact of the program. The study found that participants made substantial improvements on measures of school readiness and literacy and, by extension, indicated that the Even Start program was effective.

However, the same researchers conducted an RCT on Even Start several years later and found that although the treatment group saw improvements, these improvements were the same as those in the control group, indicating that the program had essentially no net impact. Today, the Department of

Education notes that "'pre-post' study designs often produce erroneous results," citing these evaluations of Even Start as an example.

Sources: St. Pierre et al. (1996), Ricciuti et al. (2004), and Institute of Education Sciences (2003, pp. 1–2).

Confidence in the Results

PFS stakeholders desiring a high degree of confidence in the results of a PFS project should consider an RCT, especially if they want to understand the true impact of the selected program. More than any other evaluation design, RCTs can establish the strongest causal link between a program and its effect on those who participate.[5] RCTs accomplish this because when well-designed and implemented, they essentially eliminate the risk of selection bias (the possibility that people who choose to participate will differ in some important way from those who do not) because the opportunity for program participation is not ultimately decided by individuals or program staff.

Consider, for example, an after-school program designed to engage at-risk youth. If program participation is voluntary, individuals that sign up for the program may also possess additional characteristics (e.g., motivation) that increase the likelihood they will achieve positive outcomes regardless of the program. The process of randomizing those offered the program and those not, on the other hand, creates two statistically equivalent groups in treatment and comparison, consequently producing accurate and unbiased results (Abdul Latif Jameel Poverty Action Lab, n.d.).

This addresses a practical concern for PFS stakeholders: the possibility that the project is only enrolling those participants that are most likely to succeed (a manifestation of the "creaming effect"[6]). With its strong focus on randomly allocating program slots, RCTs ensure both fairness and protect against creaming. As a result, the demonstrated results are much more likely to reflect the actual impact of a program.

Other evaluation designs use a nonrandomized comparison group and rely on quantitative methods to control for different participant background variables (e.g., income) to make the control group as similar as possible to the treatment group. In practice, however, eliminating possible biases is a significant challenge. Nonrandomized experiments have typically been shown to "yield effect estimates that either are demonstrably different or are at best of unknown accuracy" when compared with RCTs (Shadish, Clark, and Steiner 2008). For PFS projects, the results of non-RCT evaluations may not hold up as well under scrutiny.[7]

Without randomization, attributing positive outcomes to a program—rather than other reasons, including unobserved ones—is much more difficult. With randomization, one can be significantly more confident in the results.

Results with Clarity

PFS projects are high-profile, particularly at this early stage of the field's development. Many audiences (including media, government, investors, and the general public) are watching to see if these programs succeed and funders are repaid. Being able to clearly communicate the meaning (and limitations) of evaluation results helps ensure the program's impact is accurately understood across stakeholder groups.

Of all the rigorous evaluation methods available to measure success, results from RCTs provide the clearest answers about program impact. Because randomization ensures no statistical difference exists between those who received treatment and those that did not, the analytical techniques necessary to calculate program impact may be as simple as subtracting the average outcomes of the treatment group from the average outcomes of the control group. The difference between the two is an estimate of the impact of the program.

High-quality quasi-experimental designs may produce precise results, but they do so at the expense of clarity. Such designs often require significant matching techniques or other statistical corrections to measure impact.[8] This complexity makes it much harder to clearly explain the results and much easier for nontechnical audiences to misinterpret the estimated program impact. Other evaluation designs (e.g., historical baseline) that count activities or outcomes can produce simple, clear results, but they do so at the expense of confidence that they reflect true program impact. RCTs preserve confidence without sacrificing clarity by relying on the strongest possible comparison group.

Fairness of Service Delivery

Social service provision typically aims to ensure that the people who can benefit from a program actually receive it. However, rarely does a program have sufficient resources to serve every eligible applicant. Research suggests that even the most thoroughly vetted and validated programs, such as Nurse-Family Partnership and drug courts, are rarely sufficiently resourced to serve more than a fraction of the eligible target population in any one place.

Because RCTs randomize who gets service access and who does not, not all eligible people receive the service. This can sometimes be difficult for providers and government officials. But given the reality that often more people need services than a given program can provide, PFS project partners must find some method to distribute access to those scarce resources. Rather than creating an unfair distribution of resources through another method, such as first-come first-served, randomization provides all eligible people with an equal chance of receiving services. RCTs are therefore seen by some as the fairest way to determine who receives services. (Later in this brief we explore opportunities for randomization in social service delivery.)

Many critics of RCTs cite the difficulty of randomization as a barrier to using RCTs to evaluate social programs (Sullivan 2011). In practice, the work required to develop strong randomization plans can be

significant,[9] underscoring the importance of engaging a qualified evaluator at an early stage of project development. But such plans, which include developing clear criteria for eligible populations and determining the point of randomization, strengthen the program model and provide structure for consistent program implementation.

Costs and Value

Every PFS-funded program must measure success with an evaluation, and all evaluation designs have costs. Several sources estimate that evaluations should cost anywhere from 0.5 percent to 15 percent of the actual program cost.[10] These costs include the labor and expertise necessary to plan the evaluation, collaborate with stakeholders, collect and analyze data, and report results. These steps and costs are inherent to all evaluation designs, not just RCTs (Buck and McGee 2015). Because the RCT design compares a treatment group to a control group and looks at the difference in effects, the amount of work required by researchers to analyze and communicate results may be less than that needed under other quasi-experimental designs that require complex methodologies.

Approaches that simplify one or more of the steps listed above can reduce the evaluation cost for a PFS-funded intervention.[11] Two examples are assessing the independent evaluator's expertise and reducing the need for new primary data collection by relying on existing data systems.

Evaluators with strong expertise in the program and type of population served will require fewer resources during the evaluation. Evaluators experienced with RCTs and the mechanics of randomization can bring their lessons learned and existing randomization tools to a PFS project, and such evaluators are more likely to anticipate and protect against barriers to efficient program implementation and participant enrollment.

Governments with existing high-quality administrative data systems can save resources usually spent on new, primary data collection of treatment and control participants (Coalition for Evidence-Based Policy 2012). Foundations, such as the Laura and John Arnold Foundation and the Annie E. Casey Foundation, and government leadership from the White House[12] have highlighted the potential for low-cost RCTs for these localities and programs with quality data systems and large populations (Laura and John Arnold Foundation 2015).

In the context of PFS, any up-front costs for implementing a strong RCT design will likely be outweighed by the benefits accrued. Increased clarity and confidence in the results from RCTs help the government know they are funding something that works.

Building Evidence for the Intervention

Incorporating a rigorous RCT evaluation into a PFS project can serve another end: building the evidence base for a particular intervention. Using a rigorous design that supports causal conclusions builds evidence about the impact of a particular program model. Building additional evidence holds great value

because even models with the strongest research backing have usually only been tested rigorously in a few places, and that evidence might not be directly translatable to a new context.

Stronger evidence for a program's positive impact may increase the likelihood of further replication and scaling in other locations by, for example, spurring inclusion of the program in an evidence-based program clearinghouse. This type of evidence-building for a promising program can be an important ancillary goal of PFS funders and payors, but it has potentially significant implications.[13] Further, demonstrating a program's impact in a specific place and with a particular service provider may help make the case for future investment in that location.[14] Conversely, when a program does not demonstrate impact, jurisdictions can build evidence of what does not work and divert spending away from ineffective programs. Finally, collecting information after implementation on how the program was structured and delivered. This is known as a process study and helps provide context for the RCT's results.

Where Have RCTs been Incorporated into PFS projects?

RCTs are built into the design of many PFS projects in the United States across a range of outcome areas. Seven of the first 11 US PFS deals include an RCT (table 1).[15] These include programs that target prisoner reentry, foster care, housing, and health; the evaluation periods range from 4.5 to 7.5 years. The widespread application of RCTs in PFS projects across geographies and issue areas provides early verification that the design is a preferred method.

TABLE 1

Evaluation Designs of the First 11 US PFS Projects

	Evaluation design	Outcome for payment	Length of evaluation
New York City, NY NYC ABLE Project for Incarcerated Youth	Quasi-experimental regression discontinuity design with historical baseline	1. Recidivism bed days avoided	Planned as 4 years. Program ended in 3.
Salt Lake County, UT Utah High Quality Preschool Program	Longitudinal study	1. Preschool students who avoid special education placement	4-year service delivery term and 12-year repayment term and evaluation period
New York State Increasing Employment and Improving Public Safety	RCT	1. Recidivism bed days avoided 2. Indication of positive earnings after release from prison 3. Number of members who start a CEO transitional job	4-year service delivery term and 5.5-year repayment term and evaluation period
Massachusetts Juvenile Justice Pay for Success Initiative	RCT	1. Recidivism bed days avoided 2. Improved job readiness 3. Improved employment outcomes	7-year service delivery term, repayment term, and evaluation period
Chicago, IL Child-Parent Center Pay for Success Initiative	Quasi-experimental with propensity score matching	1. Decrease in special education 2. Improved job readiness 3. Improved employment outcomes	4-year service delivery term and 17-year repayment term and evaluation period
Cuyahoga County, OH Partnering for Family Success Program	RCT	1. Out-of-home foster care placement days avoided	4-year service delivery term and 5-year repayment term and evaluation period
Massachusetts Chronic Homelessness Pay for Success Initiative	Validated data (benchmark analysis)	1. Number of days participants are continuously housed[a]	5-year service delivery term, 6-year repayment term, and 5.25-year evaluation period
Santa Clara County, CA Project Welcome Home	"Intention to Treat" analysis with RCT companion study[b]	1. Number of months of stable tenancy achieved	6-year service delivery term, 6-year repayment term, and 5.25-year evaluation period
Denver, CO Social Impact Bond Program	RCT	1. Reduction in jail bed days 2. Housing stability	5-year service delivery term and repayment term and 5.25-year evaluation period

Connecticut Family Stability Project	RCT	1. Prevented out-of-home placements 2. Prevented re-referrals to DCF 3. Reduction in substance abuse 4. Family Based Recovery enrollment	4.5 years
South Carolina Nurse-Family Partnership Project	RCT	1. Reduction in pre-term births 2. Reduction in child hospitalization and emergency department usage 3. Increase in healthy spacing between births 4. Family increase in number of women served[c]	4-year service delivery term, 5-year repayment term, and 7-year evaluation period

Sources: Urban PFS Website; Nonprofit Finance Fund, "Pay for Success: The First Generation," April 2016, accessed August 8, 2016, http://www.payforsuccess.org/sites/default/files/Pay%20for%20Success_The%20First%20Generation.pdf.

Notes: CEO = Center for Employment Opportunity; DCF = Department of Children and Families; RCT = randomized controlled trial.

[a] Minimum of 12 consecutive months (with the exception of past participants whose days may count as Former Qualified Participant Days although they left the program before the 12-month mark).

[b] The RCT will not determine outcome payments.

[c] Increase of number of first-time moms served in predetermined zip codes with high concentrations of poverty.

What are the Likely Local Challenges and Potential Solutions to Developing an RCT for a PFS Project?

Standing up a PFS project with an RCT, like any new effort, can bring challenges. Before launching an RCT, PFS project development partners should closely study project and evaluation feasibility. Each PFS partner should acknowledge these challenging questions and try to find answers in discussion with the independent evaluator, an important partner that should be brought in during planning to avoid later confusion or issues.

Do We Have Enough People to Participate?

Determining the number of people eligible for a PFS-funded program is crucial for planning the evaluation planning and eventually measuring success. If there are too few eligible participants, there is a risk of concluding that the program has no impact even though a real (but undetected) benefit to participants exists.[16] Insufficient target populations are more likely in sparsely populated (e.g., rural) areas or with programs designed to help people with a very specialized need. PFS partners should work with evaluation experts to determine if an RCT is feasible for the expected size of the target population. If not, PFS teams should consider expanding the geographic boundaries of the program or determine if it makes sense to expand the eligibility criteria for receiving services. Evaluators should only proceed with the latter if it is supported by the theory of change for the proposed intervention.[17] Participants in both the treatment and the control groups of the RCT must be informed about the evaluation design and must provide their signed consent to participate. PFS planners should prepare for the possibility that some eligible participants will decline to give their consent.

How Will the Target Population Be Randomized?

PFS partners should consult with evaluators to find the best way to identify potential participants, vet their eligibility, and randomize them into treatment and control groups. Once people are identified as eligible, the evaluation staff or others outside of the service delivery organizations should randomize the assignment to minimize selection bias. In the most efficient model, program participants flow through a single organization (e.g., a single correctional facility or single school) and are randomized at that point of contact. For many programs, however, the participants may come from multiple sources and may need to be assigned.

Where Can We Find Opportunities For Randomization?

WAITLISTS AND OVERSUBSCRIPTION

Any time public need outweighs scarce resources, an opportunity exists to randomize delivery and measure impact.

In the Connecticut Family Stability pay for success project,[18] the Connecticut Department of Children and Families found that 18,118 families out of those they investigated—over 50 percent—had at least one parent who struggled with substance abuse. However, because of current resource constraints and the program's limited evidence of effectiveness, only 500 families[19] can participate in the program. If the RCT finds a positive program impact on this smaller cohort of families, it would provide evidence that supports scaling.

Random selection among even a larger population provides an opportunity to measure impact overall and for specific subpopulations. In 2010, Oregon expanded Medicaid eligibility to childless adults that met specific income criteria. Because need outweighed resources in this case as well, they randomly selected over 200,000 people to receive an invitation to apply. This included 1,827 people returning to the community from prison (Mallik-Kane et al. 2014). This randomization provided an opportunity to measure program impact for this important yet difficult-to-serve population. PFS projects designed to scale promising programs can benefit from incorporating random assignment into service delivery.

In cases such as the Connecticut PFS project and the Oregon Health Care Lottery, where the need for services clearly exceeds the operational capacity, randomization may be the most equitable means to distribute the limited available services. By choosing randomization over explicit screening criteria or a first-come first-served model, researchers serve the same number of people but limit the possibility of any bias influencing whether an individual receives services.

TESTING MULTIPLE MODELS

When faced with a choice of multiple program options, policymakers may choose to implement several different ones during a pilot phase to test which is most effective. Similarly, if many interventions seem promising for a particular problem, researchers can implement several of them and randomize between the various interventions and a control group. Such a design allows researchers to test the relative merits of each intervention against the status quo. For example, HUD's Family Options Study[20] implemented a multisite RCT comparing outcomes for four discreet interventions: housing subsidy only, project-based transitional housing, community-based rapid rehousing, and existing care (control group).[21]

Though the PFS field has not yet implemented an RCT in this way, a PFS project could use a similar approach if several distinct interventions have demonstrated results. The project could select the same outcome for all of the interventions and pay based on that outcome. Such a study would demonstrate not just the effectiveness of the program compared with a control group but also the effectiveness of each intervention relative to the others (i.e., which of several approaches is most effective).

Once Randomized, How Will We Know What Happens to Participants?

RCT data analysis produces an estimate of program impact. That analysis is only as good as the data that support it.

PFS partners will need a clear understanding of existing local data systems, the reliability of the data in those systems, and their ability to individualize data regardless of a person's group assignment.[22] If existing data systems are insufficient, PFS partners need to figure out if new systems can be implemented during and after program participation to track outcomes for people in both the treatment and control groups. After randomization, it is critical to track outcomes for people in both experimental groups on the key outcomes of interest. The ability to draw a conclusion about the effectiveness of the program is dependent on the ability of the study to track and analyze information about the people in both the treatment and the control groups. PFS partners considering an RCT should take care to clearly identify how they will define the treatment group.[23]

How Can We Incorporate Intermediate Success Measures?

Using an RCT design to measure program impact in a PFS project does not preclude using additional data to determine intermediate metrics of success. Projects can pair a formal RCT evaluation with process measures, such as the number of clients who successfully complete a program, that allow intermediate success payments before the completion of the overall evaluation. This approach has already taken hold in some existing projects.[24]

How Can We Build Support in the Community?

Building support for an RCT extends beyond the leaders creating a PFS project. Gaining the buy-in of community leaders, service providers, and other stakeholders is often also important. In particular, program staff or the target population may be resistant to a design that may be perceived as limiting the number of people able to access the service. To address this concern, organizers should make the case that the project's success has real value for participants and service providers and that an RCT is a fair way to distribute scarce resources while providing a strong way to measure the program's effectiveness and determine whether it should be scaled (Gueron 2002). Providing venues to explain the benefits of the project and the evaluation design, answer questions, and acknowledge challenges in a transparent manner are critical steps to ensuring long-term success.

It is also important to make the randomization process as easy as possible for the staff and service providers. PFS project leaders should seek to minimize disruption within the participant intake process (Gueron 2002). In several early PFS projects, this has involved partnering with qualified evaluators to run the randomization process Cunningham et al. (2016).

Conclusions

Partners in PFS projects must measure success to make key decisions both during and after program delivery. The most rigorous and accurate way to do that is with an RCT. Although not without challenges and costs, RCTs are usually the best way to confidently answer the question, "Did this program work?"

Given the strengths of the model, planning an evaluation for any PFS project should start with a consideration of RCTs: Are they appropriate for this program? Do we have the population and data systems in place to support one? And if not, how can we fix those gaps? How can we secure community buy-in?

Results of an RCT offer the most clarity and confidence in the final outcomes, empowering PFS stakeholders to communicate those results accurately to the media and the community. They also preserve fairness in the face of scarce resources by offering eligible participants an equal opportunity to receive services. In this way, using an RCT can be a miniature model of a PFS transaction itself: up-front investment work on evaluation design provides a better long-term payoff. Stakeholders (such as funders, governments, service providers, and the public) benefit from this simplicity because the results are easier to understand than with other models.

Advocates of RCTs must acknowledge and have a plan to remediate community-specific challenges that threaten feasibility. For these reasons, evaluators must recognize the value and insights from community leaders during planning stages, and community leaders must rely on skilled evaluators to answer tough questions about sample size, eligibility determination, and the need to preserve the random assignment.

As the ultimate arbiter of success, PFS projects must include a strong evaluation. Different parties will want as much confidence as possible in the impact of the specific program in a PFS project. Many evaluation approaches exist with varying degrees of rigor, transparency, and objectivity. Out of these available evaluation designs, RCTs are widely considered the strongest option. In general, it is wise to heed the words of William Shadish, coauthor of the seminal text book on evaluation design: "If you can do a randomized trial, by all means do it," (Clay 2010).

Notes

1. For example, RCTs are seen as providing the strongest evidence by a variety of evidence-based program clearinghouses, including Blueprints for Healthy Youth Development ("Program Criteria," accessed August 8, 2016, http://www.blueprintsprograms.com/criteria), the Department of Education's What Works Clearinghouse (Institute of Education Sciences 2011), and the National Registry of Evidence Based Programs and Practices on substance abuse and mental health ("Program Review Criteria," accessed August 8, 2016, http://www.nrepp.samhsa.gov/04e_reviews_program.aspx).
2. For more information on RCTs, see "Data & Methods: Experiments," Urban Institute, accessed August 4, 2016, http://www.urban.org/research/data-methods/data-analysis/quantitative-data-analysis/impact-analysis/experiments
3. We're using a broad definition of "program" because RCTs can be used to test a wide range of things (e.g., messaging).
4. When sites and target populations are chosen purposefully through a nonrandom process, they are not necessarily representative of the broader population eligible for the program. In other words, evidence gathered through an RCT may have limited "external validity" (Olsen et al. 2013). That said, there are methods by which researchers can select the sample that improve this generalizability (see, for example, a recently proposed approach from Tipton et al. [2014]).
5. Even though RCTs can clearly and confidently answer "Did it work?," they do not reveal *why*. For example, an RCT may show that a workforce development program successfully increased the likelihood that participants

found full-time work. But it cannot prove why this happened or whether a certain part of the program (e.g., interview training or skills development) was more effective than others.

6. As in: selecting the best of something; the cream of the crop.
7. The PFS project focused on early education in Utah provides an illustrative example. To evaluate the impact of an early education program, researchers identified a group of 120 children that they thought would end up in special education without additional services and offered participation in a preschool program. At the end of the first year, they found that in fact 119 out of the 120 children did not end up in special education when entering kindergarten. Although this result seemed especially promising to many observers, it also attracted significant skepticism from critics. Because the program had not established a comparison group through an RCT (or another design), there was no way to know if the program was causing such dramatic impacts for the population or if there were other reasons (e.g., the process of identifying children at-risk of special education placement was too broad).
8. Such as propensity score matching, Heckman corrections, difference-in-difference, and regression discontinuity.
9. For example, see the description of the referral and randomization strategy used in the Denver PFS project (Cunningham et al. 2016, 9–13).
10. Molly Engle, "How Much Should an Evaluation Cost," *Evaluation is an Everyday Activity* (blog), July 24, 2012, http://blogs.oregonstate.edu/programevaluation/2012/07/24/how-much-should-an-evaluation-cost/
11. For instance, randomization simplifies the analysis necessary to calculate impact. Other evaluation designs that require more advanced statistical matching may have higher data analysis costs than RCTs.
12. Maya Shankar, "How Low-Cost Randomized Controlled Trials Can Drive Effective Social Spending," WhiteHouse.gov blog, July 30, 2014, https://www.whitehouse.gov/blog/2014/07/30/how-low-cost-randomized-controlled-trials-can-drive-effective-social-spending.
13. "Improving Outcomes through Pay for Success," Whitehouse.gov, accessed August 8, 2016, https://www.whitehouse.gov/sites/default/files/omb/budget/fy2016/assets/fact_sheets/improving-outcomes-through-pay-for-success.pdf.
14. Sam Schaeffer, Jeff Shumway, and Caitlin Reimers Brumme, "After Pay for Success: Doubling Down on What Works," *Stanford Social Innovation Review*, August 20, 2015, http://ssir.org/articles/entry/after_pay_for_success_doubling_down_on_what_works
15. Six projects use an RCT as their primary evaluation design while one (in Santa Clara, CA) uses an RCT as a secondary evaluation to determine program effectiveness but not outcome payment.
16. If you flip a coin, probability tells us that you should expect to get heads 50 percent of the time. However, it's possible that if you flip the coin 10 times, you could get fewer (or more) than 5 because chance because your sample is small. With a greater sample size (e.g., 100), you're more likely to get a result closer to that true probability.
17. Defined by Tatian (2016) as "a conceptual road map that outlines how a series of actions can bring about a desired outcome."
18. "Connecticut Family Stability Pay for Success Project," PayforSuccess.org, accessed August 8, 2016, http://www.payforsuccess.org/sites/default/files/CT-Family-Stability-PFS_Fact-Sheet_vFINAL.pdf.
19. "Connecticut Family Stability Project," Urban Institute, accessed August 8, 2016, http://pfs.urban.org/pfs-project-fact-sheets/content/connecticut-family-stability-project.
20. "The Family Options Study," US Department of Housing and Urban Development, Office of Policy Development and Research, accessed August 8, 2016, https://www.huduser.gov/portal/family_options_study.html.
21. Ibid.
22. Note that individualized data do not eliminate the ability to also anonymize data and safeguard individual privacy. For example, individuals in the treatment and control groups can be given numerical identifiers after randomization that are linked to their outcomes but not to any identifying demographic characteristics.

23. For example, in an evaluation, analysis can examine the impact of a program on a treatment group that includes all people offered the program (regardless of their actual participation) or only people who actually participated in the program. Evaluators call the first model "intent to treat" and the second "treatment on the treated," and both have advantages and limitations.

24. For example, the Recidivism Reduction and Employment project uses an RCT to evaluate outcomes, which include recidivism bed days avoided, improved job readiness, and improved employment outcomes. The project also includes interim success payments of $789 for each Roca participant who engages with a Roca youth worker nine or more times in a quarter.

References

Abdul Latif Jameel Poverty Action Lab. n.d. *Introduction to Evaluations*. Cambridge, MA: Massachusetts Institute of Technology. https://www.povertyactionlab.org/sites/default/files/resources/Introduction%20to%20Evaluations%20%281%29.pdf.

Buck, Stuart, and Josh McGee. 2015. *Why Government Needs More Randomized Controlled Trials: Refuting the Myths*. Houston, TX: Laura and John Arnold Foundation. http://www.arnoldfoundation.org/wp-content/uploads/2015/07/RCT_FINAL.pdf.

Clay, Rebecca A. 2010. "More Than One Way to Measure." *Monitor on Psychology* 41 (8): 52. http://www.apa.org/monitor/2010/09/trials.aspx.

Coalition for Evidence-Based Policy. 2012. Rigorous Program Evaluations on a Budget: How Low-Cost Randomized Controlled Trials Are Possible in Many Areas of Social Policy. Washington, DC: Coalition for Evidence-Based Policy. http://www.payforsuccess.org/sites/default/files/rigorous-program-evaluations-on-a-budget-march-2012.pdf.

Cunningham, Mary, Mike Pergamit, Sarah Gillespie, Devlin Hanson, and Shiva Kooragayala. 2016. *Denver Supportive Housing Social Impact Bond Initiative.* Washington, DC: Urban Institute. http://www.policyinnovationlab.org/wp-content/uploads/2016/02/SIC-Denver-Supportive-Housing-Social-Impact-Bond-Initiative-Evaluation-and-Research-Design.pdf.

Gueron, Judith M. 2002. "The Politics of Random Assignment: Implementing Studies and Impacting Policy." In *Evidence Matters: Randomized Trials in Education Research*, edited by Frederick Mosteller and Robert Boruch. Washington, DC: Brookings Institution Press. https://www.brookings.edu/book/evidence-matters/.

Haskins, Ron, and Jon Baron. 2011. *Building the Connection between Policy and Evidence: The Obama Evidence-Based Initiatives.* London: Nesta. http://www.brookings.edu/~/media/research/files/reports/2011/9/07-evidence-based-policy-haskins/0907_evidence_based_policy_haskins.pdf.

Institute of Education Sciences. 2003. *Identifying and Implementing Educational Practices Supported by Rigorous Evidence: A User Friendly Guide.* Washington, DC: US Department of Education. http://www2.ed.gov/rschstat/research/pubs/rigorousevid/rigorousevid.pdf.

———. 2011. What Works Clearinghouse: Procedures and Standards Handbook (Version 2.1). Washington, DC: US Department of Education.

Laura and John Arnold Foundation. 2015. *Request for Proposals Low-Cost Randomized Controlled Trials to Drive Effective Social Spending*. Houston, TX: Laura and John Arnold Foundation. http://www.arnoldfoundation.org/wp-content/uploads/Request-for-Proposals-Low-Cost-RCT-FINAL.pdf.

Mallik-Kane, Kamala, Akiva Liberman, Lisa Dubay, and Jesse Jannetta. 2014. *Prison Inmates' Prerelease Application for Medicaid: Take-Up Rates in Oregon.* Washington, DC: Urban Institute. http://www.urban.org/sites/default/files/alfresco/publication-pdfs/413199-Prison-Inmates-Prerelease-Application-for-Medicaid.PDF.

Olsen, Robert B., Larry L. Orr, Stephen H. Bell, and Elizabeth A. Stuart. 2013. "External Validity in Policy Evaluations That Choose Sites Purposively." *Journal of Policy Analysis and Management* 32 (1): 107–21. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4138511/.

Ricciuti, Anne E., Robert G. St. Pierre, Wang Lee, and Amandra Parsad. 2004. *Third National Even Start Evaluation: Follow-Up Findings from the Experimental Design Study*. Washington, DC: US Department of Education. http://ies.ed.gov/ncee/pdf/20053002.pdf.

St. Pierre, Robert G., Janet P. Swartz, Stephen Murray, and Dennis Deck. 1996. *Improving Family Literacy: Findings from the National Even Start Evaluation*. Bethesda, MD: Abt Associates Inc. http://www.abtassoc.us/reports/paper5.pdf.

Shadish, William R., M. H. Clark, and Peter M. Steiner. 2008. "Can Nonrandomized Experiments Yield Accurate Answers? A Randomized Experiment Comparing Random and Nonrandom Assignments." *Journal of the American Statistical Association* 103 (484). http://stat-athens.aueb.gr/~jpan/Shadish-JASA2008(1334-1356)-17mr09.pdf.

Sullivan, Gail M. 2011. "Getting Off the 'Gold Standard': Randomized Controlled Trials and Education Research." *Journal of Graduate Medicine Education* 3 (3): 285–89. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3179209/.

Tatian, Peter A. 2016. "Performance Measurement to Evaluation." Washington, DC: Urban Institute. http://www.urban.org/research/publication/performance-measurement-evaluation-0.

Tipton, Elizabeth, Larry Hedges, Michael Vaden-Kiernan, Geoffrey Borman, Kate Sullivan, and Sarah Caverly. 2014. "Sample Selection in Randomized Experiments: A New Method Using Propensity Score Stratified Sampling." *Journal of Research on Educational Effectiveness* 7 (1): 114–35. http://www.tandfonline.com/doi/abs/10.1080/19345747.2013.831154.

Theodos, Brett, Margaret Simms, Rachel Brash, Claudia Sharygin, and Dina Emam. 2014. "Randomized Controlled Trials and Financial Capability: Why, When, and How." Washington, DC: Urban Institute. http://www.urban.org/sites/default/files/alfresco/publication-pdfs/413172-Randomized-Controlled-Trials-and-Financial-Capability.PDF.

About the Authors

Justin Milner is director of the Pay for Success Initiative and a senior research associate in Urban Institute's Policy Advisory Group. His work focuses on the intersection of research, policy, and practice; supporting efforts to engage effectively with policymakers and practitioners in the application of research findings; and the development of new evidence. His past experience includes roles at the Annie E. Casey Foundation and the US Department of Health and Human Services. He received a BA in political science from Yale University and an MPA from the Woodrow Wilson School at Princeton University.

Kelly Walsh is a senior research associate in the Policy Advisory Group and the Justice Policy Center at the Urban Institute and managing director of the Pay for Success Initiative. She is most interested in the efficacy of forensic processes, the causes of wrongful convictions, and the mechanisms of private investment for the sake of public good. Before joining Urban, Walsh was an instructor at the John Jay College of Criminal Justice and a researcher at the Center for Modern Forensic Practice. She earned a BS in chemistry from the University of Scranton and a PhD in criminal justice, with a specialization in forensic science, from the Graduate Center of the City University of New York.

Acknowledgments

This brief was funded by the Laura and John Arnold Foundation. We thank our funders, who make it possible for Urban to advance its mission.

The views expressed are those of the authors and should not be attributed to the Urban Institute, its trustees, or its funders. Funders do not determine research findings or the insights and recommendations of Urban experts. Further information on the Urban Institute's funding principles is available at www.urban.org/support.

The authors thank Mary Cunningham and Brett Theodos for commenting on an earlier version of this paper.

2100 M Street NW
Washington, DC 20037

www.urban.org

ABOUT THE URBAN INSTITUTE

The nonprofit Urban Institute is dedicated to elevating the debate on social and economic policy. For nearly five decades, Urban scholars have conducted research and offered evidence-based solutions that improve lives and strengthen communities across a rapidly urbanizing world. Their objective research helps expand opportunities for all, reduce hardship among the most vulnerable, and strengthen the effectiveness of the public sector.