An Introduction to Evaluation Designs in Pay for Success Projects

Kelly A. Walsh, Rebecca TeKolste, Ben Holston, and John K. Roman

September 2016

Evaluations help discern the links between program activities and their consequences. Did participating in a drug treatment program increase the participants' chances of maintaining sobriety? Did attending preschool increase the likelihood of students' future academic success? Most social programs are planned and implemented without the ability to confidently answer these types of questions. But programs funded through pay for success (PFS) partnerships require these answers and, in doing so, require an independent evaluation that observes outcomes and estimates impacts.

In PFS projects, evaluations trigger payments to funders by determining if the project's outcome targets are met or exceeded. By making payment dependent on achievement, the PFS structure encourages governments to focus on outcomes and measure success through rigorous, transparent, and objective evaluations.

For stakeholders pursuing PFS projects, however, considering how best to incorporate evaluation into project design can be confusing and even intimidating. This brief provides a basic overview of evaluation designs to assist PFS stakeholders engaged in deal development.[1] It focuses on the concept of comparison and its relation to various designs, and it presents key questions that PFS planners should address as they participate in evaluation design discussions.

Three accompanying briefs look specifically at broader lessons from the PFS evaluation field on the value of randomized controlled trials for PFS evaluations (box 1).

BOX 1

Other Urban Institute Briefs Exploring Evaluation and Pay for Success

- "Measuring Success in Pay for Success: Randomized Controlled Trials as the Starting Point" (Milner and Walsh 2016) explores in greater depth the value of randomized controlled trials in the context of PFS projects and considers potential opportunities, challenges, and solutions related to their use.
- "Core Components of a Pay for Success Research Design: Lessons Learned" focuses on the role of evaluation partners during the early stages of PFS feasibility analysis and structuring and includes lessons from navigating the unique dynamics and requirements of the PFS research design process. The Urban Institute plans to publish this brief sometime in late 2016.
- "Practical Considerations for Pay for Success Evaluations" reviews PFS evaluation experiences, helping to pave a more established path forward for future PFS evaluators. The Urban Institute plans to publish this brief sometime in late 2016.

Evaluation Designs

There are a range of evaluation designs available to PFS project administrators, each with a different level of analytical rigor and with a different ability to measure (or approximate) causation.

A randomized controlled trial (RCT), or *experiment*, randomly assigns eligible people to a treatment group, which receives services, or to a control group, which does not. As a result, the differences reflect only the direct effect of the program. The RCT design accounts for all other competing explanations. This method produces causal results with the highest confidence and clarity.

Designs using difference-in-differences, paired testing, and regression discontinuity are examples of *quasi-experimental* evaluation designs. Most compare outcomes of a group of people that received services to a group of people that did not receive services, although the two may differ in meaningful ways (and thus bias the results). These designs attempt to minimize potential bias from competing explanations (for example, the treatment group might be older or more motivated on average than the comparison group). These designs use various statistical methods to approximate the rigor of the RCT. When compared to nonexperimental methods, these designs produce results with fewer biases and increase confidence that the treatment was causally related to better outcomes.

Pre-post analysis and benchmark designs[2] are examples of nonexperimental evaluations because neither includes an untreated comparison or control group. Because these designs do not compare the outcomes of the people treated by the program to those of the people not treated, these designs may lead to conclusions that a more sophisticated design might contradict.

Evaluations are built into the design of all US PFS projects across a wide range of outcome areas. Seven of the first 11 US PFS projects include an experimental evaluation design (appendix A).[3] Two projects, in New York City and Chicago, include a quasi-experimental design (QED), and another two,

the Utah High Quality Preschool Program and the Massachusetts Chronic Individual Homelessness PFS project, rely on nonexperimental evaluation designs to measure success and trigger repayment.

Regardless of the design, any evaluation that strives to measure program impact must choose outcomes to observe and measure (box 2) and compare those outcomes to an untreated but equivalent comparison group. To determine whether observed outcomes of the treatment group were *caused* by the program, we must ask how those outcomes fared compared to expected outcomes absent program services. The best program evaluations therefore seek to create a comparison group composed of individuals who could have been eligible for the program, but owing to capacity, time, or geography, for instance, were not offered the program and did not receive the treatment. This comparison approximates how the treated group would have performed without treatment. The different ways that evaluation designs create the comparison group, and their relative success in doing so, is the primary distinguishing factor between these designs.

BOX 2

Outcomes and the Theory of Change

PFS projects typically choose between one and three outcomes to determine repayment.[a] Harrell (1996) identifies four factors to consider when choosing outcomes:

- **Relevance:** Do outcomes reflect the effect of services that will actually be delivered and align with the existing program theory of change? Do they capture the priorities of all PFS stakeholders?
- **Comprehensiveness:** Are the outcomes paired with other performance measures that cover inputs, outputs, and service?
- **Extent of program control:** Does the program have influence or control over outcomes measured? If control is limited, the outcomes may not fairly reflect program success.
- **Validity:** Are the outcomes of interest supported by the available data and free from reporting bias for both the treatment and comparison groups?

An established theory of change[b]—a framework that describes "an organization's work that links inputs, activities, outputs and outcomes"—helps ensure that the four factors above are addressed in the evaluation planning stage (Winkler, Theodos, and Grosz 2009, 4). Key questions for PFS planners include the following:

- Does the program have an existing theory of change?
- Can that theory be updated to reflect the participants' specific geography?
- Are the outcomes that trigger repayment reflected in the model?

[a] PFS Project Fact Sheets," Urban Institute website, http://pfs.urban.org/pfs-project-fact-sheets.
[b] This is also known as a logic model.

KEY QUESTIONS AN EVALUATION DESIGN SHOULD ADDRESS

PFS planners engaged in evaluation planning and design should address the following:

- Should the performance of the treatment recipients be compared against a similar group of nonrecipients?
- How is this comparison constructed (e.g., matched group or random assignment)?
- How will different points of comparison (e.g., a person's history before receiving services, matched comparison group, or randomized control group) affect the ability to actually measure success?

Experimental Design

RCTs are experiments designed to answer a single question: Will people (or classrooms, communities, or other units of interest) who receive program services experience greater beneficial outcomes than those who do not?

RCTs are "generally considered the most reliable way to determine a program's impact" (Tatian 2016, 9).[4] As described in our companion brief, "Measuring Success in Pay for Success: Randomized Controlled Trials as the Starting Point," "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" (Milner and Walsh 2016, 2).

Milner and Walsh (2016, 2–3) describe this method:

> 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.

The advantages of an RCT design include the following:

- The highest confidence and clarity in the results among common social science designs
- A mechanism to enable fair distribution of scarce resources
- Contribution to the program's evidence base

KEY QUESTIONS IF CONSIDERING AN RCT

All PFS evaluation planning efforts should begin by considering an experimental design. Planners considering an RCT should address these key questions early in project planning:

- Do we have enough people to participate in both the treatment and control groups?[5]
- How will the target population be randomized, and who will control the process?
- Is this program suitable for randomization?[6]
- Once the population is randomized, can we measure their service receipt and, equivalently, measure outcomes for both treatment and control?
- How can we build support in the community for randomization?

For guidance on these questions, and for an in-depth discussion on why PFS planners should choose RCT evaluation designs, see the companion brief by Milner and Walsh (2016). Even if an RCT is not ultimately selected, the planning process itself can provide critical lessons learned for the design of a high-quality quasi-experiment.

Quasi-Experimental Designs

Many QEDs exist, making this class of evaluations difficult to generalize. However, QEDs do share a unifying feature: they seek to create a comparison group that is as identical as possible to the treatment group in the absence of randomization. QEDs can take myriad approaches to create this nonrandomized comparison group and analyze the results.

The best QEDs seek to approximate an RCT through a variety of statistical controls. The Urban Institute has created an extensive resource on the various designs.[7] In this list, we briefly summarize some of the common designs.

- **Regression analysis:** Regressions use statistical analysis to separate preexisting differences in the treatment and the comparison groups that might provide alternative explanations for the ultimate outcome. Evaluations using regressions acknowledge that participant outcomes may vary for many reasons, only one of which is the program. The goal of a regression analysis is to isolate the program impact by controlling for other factors that might affect the outcome.
- **Regression discontinuity design (RDD):** A variant of regression analysis, an RDD compares two groups: one directly above an eligibility threshold and another directly below. Participants above the eligibility threshold receive the program while those below do not. Because these two groups are so similar before the application of the program, any differences measured after the program can generally be attributed to the program. Education and behavioral social services implement RDDs frequently because many programs select based on eligibility. If students who barely made the cutoff improve markedly compared with students who just missed it, then the program can be considered positive. However, because many individuals may not always cluster around a threshold, researchers can select this method in specific circumstances only. The need to increase sample size leads to enrollment of individuals further from the threshold, and thus the validity of the method diminishes.

- **Propensity scores:** Generally, comparison and treatment groups could be expected to have different outcomes. But outcomes may be affected by important but unobservable factors that are not included in the analysis because they cannot be measured. In these situations, researchers sometimes rebalance the data to make the two groups look more like the observations when the treatment and comparison groups consistently overlap. The approach is essentially a weighting mechanism that makes the individuals who did not receive a program look statistically similar to the people who did, and vice versa. The method uses all of the available observable data about the treatment and comparison groups and weights those attributes so that the two groups appear identical, thus mimicking an RCT. The approach has strong merits in the presence of sufficiently large sample sizes, a broad range of variables, and a lack of attrition. However, if important individual attributes related directly to success are unobservable, such as motivation, this approach has more value as a diagnostic tool and can lead to inaccurate conclusions if overinterpreted.

- **Difference-in-differences:** This approach measures the change in outcomes experienced by the treatment group (receiving the program) from before project launch to project completion, or the treatment group's *difference* over time. The comparison group's difference over time is similarly measured. The two differences are compared, and the difference-in-differences is interpreted as the program's estimated impact. For example, if the treatment group experienced a 10 percent improvement in test scores over a defined period while the comparison group experienced a 5 percent increase, then, if other factors are controlled for, the program produced twice as much improvement as business-as-usual service delivery.

- **Instrumental variables:** An instrumental variable helps a statistical equation better represent the relationship between a causal variable and the outcome. In general, a linear regression assumes that no correlation exists between the causal variable and the error term (the amount the predicted outcome differs from the real outcome). However, when a correlation exists between the causal variable and the error term, we might be able to find another variable that has a relationship with that causal variable but *no* relationship with the error term. The second variable, the instrumental variable, will help us isolate the effect of the first variable on the outcome.

These designs can control for several factors that could bias the results (e.g., race, gender, income level, education, motivation). Without randomization, however, it is not possible to account for all factors, including unobservable ones, that might affect the outcomes and introduce bias. These designs use all data available for *both* the treatment and comparison groups that might affect an outcome to attempt to balance (make as similar as possible) the treatment and comparison groups. Notably, if data are only available for the treatment group, those data cannot be used in any of the designs described previously (box 3).

For example, many evaluators study programs that are voluntary. The willingness to volunteer might be indicative of other nondemographic characteristics (e.g., motivation) that could bias participation toward favorable outcomes. A quasi-experimental design that simply compares the

treatment group that volunteered with a group with similar demographic characteristics that refused the program would bias results toward a larger-than-true effect. An RCT that randomizes volunteers into treatment and comparison groups controls for the effect of motivation and other observable and unobservable traits (e.g., native ability, intrinsic perseverance, patience) that may affect a person's likelihood of achieving success regardless of the PFS-funded program.

BOX 3

Quality and Scope of Data

Individual-level data for people who receive program services and those who do not are critical to any evaluation design. PFS planners should address these queries:

- Do data already exist to establish a baseline for characteristics of the target population and the comparison group that are expected to affect their outcomes? Can these data be collected before the intervention?
- Is collecting data feasible for the planned outcomes for both groups?
- Are different data collection and management strategies required to collect data from intervention participants and the comparison group?

Nonexperimental Designs

Nonexperimental evaluation designs either do not include a comparison or include a weak one that does not control for biases. These designs measure outcomes for only those served by a program. Success is determined either by comparing those outcomes to some predetermined benchmark from a larger population or by comparing treatment group outcomes to expected performance absent the intervention. These types of design generally do not control for any of the myriad other factors that may influence outcomes.

Some PFS projects, such as the Massachusetts Chronic Individual Homelessness PFS project, have used a benchmark to determine success payments, rather than a design with a comparison group. Originally designed for social impact bonds in the United Kingdom, these benchmark designs—also known as *rate card* programs—define benchmarks for only treatment group outcomes, which later trigger success payments. In these designs, if the treatment group meets the target benchmark, the government pays the investor.

There is no comparison group to determine whether the outcomes achieved were actually caused by the program or would have occurred even in the program's absence. For example, in a workforce development program for people recently released from prison, a rate card might have a $200 repayment for each week that a formerly incarcerated person is employed post release. Using a

benchmark based on historical data, the government predicts the likely outcomes for the target population (in the absence of the program) and establishes the rate for payment when those outcomes are met or exceeded.

However, a rate card design cannot adjust for the impacts of broader trends. Other macroeconomic changes, such as an increase in overall employment opportunities for all in the target geography, could have increased the likelihood that program participants would have found work, rather than the intervention itself affecting work. Without a point of comparison, a benchmark analysis design says little about whether the intervention actually caused the outcome it intended.

Common nonexperimental designs include the following:

- **Qualitative analysis:** Qualitative data analysis refers to many methods of data gathering and analysis that do not involve researcher control. Whether generated by observation, interviews, analysis of primary documents, site visits, or other methods, qualitative data by its nature cannot control for factors other than the program (Gibbs 2007, x).
- **Pre-post analysis:** This method uses a single individual or unit as its own comparison by charting changes over time (or over another dimension, such as how the individual changes in a different setting). A pre-post design may misrepresent the effect of the treatment when changes that would naturally occur over time are incorrectly attributed to the treatment.
- **Benchmark comparison:** Established benchmarks such as the national averages for a standardized test are often used as the basis for comparison. The Chicago Child-Parent pay for success project used national standards to determine repayment without a comparison group (Gaylor et al. 2016). Benchmarks do not control for characteristics of treatment participants, and to the extent that those are different from national averages, the results will be biased.
- **Rate cards or no comparison group:** The United Kingdom has developed an approach to evaluation that does not include a comparison group.[8] Instead, it pays a given rate directly for an outcome achieved (e.g., improved attendance in school). By establishing a rate that it is willing to pay for an outcome without comparison, the United Kingdom is not incentivizing researchers to develop an understanding of how the outcome was achieved. With no comparison group, the researcher and practitioner can glean very little understanding of the effect of the program on the outcome.

Although nonexperimental designs offer some attractive benefits for governments,[9] including simpler evaluation administration, they have a significant flaw. Of the three models discussed in this brief, nonexperimental designs produce results with the lowest confidence and cannot account for any biases that may have affected the outcomes measured. This inadequacy is particularly important for governments that want to know whether a program was actually effective and whether it should be scaled.

KEY QUESTIONS IF CONSIDERING A NONEXPERIMENTAL DESIGN

If PFS planners are considering a nonexperimental design they should answer the following:

- Is a more rigorous evaluation design feasible in this context?
- Does the design have anything against which to compare the outcomes measured?
- Is an understanding of the program's effect important or is a quick and clear measurement of outcomes achieved sufficient?

Conclusions

PFS provides governments a means to invest in evidence-based programs and to measure the impact of that investment through an independent evaluation. Strong evaluation designs will always require a point of comparison, and the best designs create that comparison through random assignment. If PFS partners plan on another method, they must understand the inherent limitations in the final results and ensure that all stakeholders, particularly funders, accept those limitations (appendix B).

This brief presents a series of key questions that PFS planners should address early in project planning. The following are critical inquiries that should inform early PFS planning discussions:

- Are there target outcomes that can be observed?
- Can a comparison group be created?
- Can outcomes be tracked for both the treatment and a comparison group?

A randomized controlled trial offers a PFS project the best chance of minimizing any bias in the results by creating the control group that most closely resembles the treatment group. This method produces the most rigorous and objective estimate of program impact. The goal of an evaluation with a PFS project structure should be either to implement or to approximate an RCT to every extent possible.

Understanding the primary considerations of a strong evaluation design will not replace the role of an expert independent evaluator in a PFS project. Nor is ensuring evaluation quality wholly sufficient. However, building an understanding across all PFS project planners to ensure accurate comprehension of any proposed designs and transparency around the advantages and limitations is critical. A strong design is essential to allowing stakeholders to anticipate future program outcomes during the deal-structuring phase and to interpret a program's outcomes with confidence. In turn, this confidence is critical for PFS projects and for broader evidence-informed government decisionmaking.

Appendix A. Evaluation Designs of the First 11 US PFS Projects

Project	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	Program planned to last 4 years but ended in 3 years
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 Recidivism and Workforce Development Project	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 outcome	4-year service delivery term and 17-year repayment term and evaluation period
Cuyahoga County, OH Partnering for Family Success Program	RCT	4. Out-of-home foster care placement days avoided	4-year service delivery term and 5-year repayment term and evaluation period
Massachusetts Chronic Individual 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. Prevention of out-of-home placements 2. Prevention of 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 preterm births 2. Reduction in child hospitalization and emergency department use 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: The Urban Institute, Pay for Success Website, http://pfs.urban.org/; Nonprofit Finance Fund (2016).

Note: CEO = Center for Employment Opportunity; DCF = Department of Children and Families; PFS = pay for success; RCT = randomized control trial.

[a] Participants are continuously housed for a minimum of 12 consecutive months (with the exception of past participants whose days may count as former qualified participant days even though they left the program before the 12-month mark).

[b] The RCT will not determine outcome payments.

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

Appendix B. How Different Design Types Create Comparisons

Consider a scenario in which a county would like to test the effectiveness of a program aimed at improving the job prospects of long-term unemployed individuals through a combination of technical skills training and interpersonal communication courses. The following table shows how each of the three types of evaluation designs might create a comparison and measure the program's impact.

TABLE B.1

Approaches to Creating a Comparison

Design	Sample approach
Experimental (i.e., randomized controlled trial)	The program is implemented in Town A in the county, and eligible residents interested in the program are randomly assigned to treatment or control groups. Because residents have been randomly assigned and because the sample size is large enough, the treatment and control groups each comprise two presumably identical groups of people. The program services are implemented, and at the end of implementation, outcomes achieved are measured for both groups. The outcomes for the control group represent the business-as-usual baseline, the outcomes for the treatment group represent the effects of the program, and the difference between the outcomes measures the program's impact (Milner and Walsh 2016).
Quasi-experimental	The same program is implemented in Town A, and residents are chosen using an assessment diagnostic or other screening process. Selected residents receive the treatment—the treatment group. Residents interested in and eligible for the program who volunteer too late are not treated, but their outcomes will be tracked anyway—the comparison group. Researchers in this scenario decide to use a difference-in-differences approach—a type of quasi-experimental design. The performance of the treatment group (difference from program start to finish while the group is receiving treatment) is compared against the performance of the comparison group (difference over the same time period with the group receiving business-as-usual services only). The difference between the treatment and comparison groups' performances is meant to estimate impact, although undetected differences in the two groups, rather than the program itself, possibly are responsible for some, most, or all of the difference in performance. Therefore, although quasi-experimental designs, including this one, help show the program's causal impact, they are not immune to unobserved variables and biases. Thus, one can have less confidence in this causal link.
Nonexperimental	Participants in Town A are chosen to participate in the program on the basis of some screening process—the treatment group. In the pre-post design, the performance of participants in the treatment group is compared to their own baseline state before treatment. However, this comparison is weak because other factors may have affected the entire population over this time period. For example, although unemployment among program participants may have decreased 20 percent during the course of the project, it may have decreased that much among the broader population as well. Therefore, this design offers limited ability to establish a causal link with a degree of confidence. In the benchmark or rate card approach, the treatment group's outcomes are simply compared to national norms or stakeholder expectations about a successful program's achievements.

Notes

1. Readers who lack formal evaluation expertise should strongly consider engaging an evaluation expert at the beginning of the project to advise on design.
2. Defined in the Nonexperimental Designs section.
3. Six projects use RCT as their primary evaluation design while one (in Santa Clara County, CA) uses an RCT as a secondary evaluation to determine program effectiveness but not outcome payment.
4. Because RCTs control for competing explanations through the design rather than through control variables, they require fewer observations (smaller sample sizes) than do quasi-experimental designs.
5. For more information on various quasi-experimental designs, visit http://www.urban.org/research/data-methods/data-analysis/quantitative-data-analysis/impact-analysis/quasi-experimental-methods.
6. Answering this question requires knowledge about, for example, whether the program is effective (if so, it may be unethical to deny treatment) or whether resources are so scarce that providing the service to all eligible individuals is not possible (i.e., some individuals will be denied service anyway).
7. For more information on various quasi-experimental designs, visit http://www.urban.org/research/data-methods/data-analysis/quantitative-data-analysis/impact-analysis/quasi-experimental-methods.
8. Matthew Eldridge, "How the U.K. Pays for Success," *PFS Perspectives* (blog), the Urban Institute, May 23, 2016, http://pfs.urban.org/pay-success/pfs-perspectives/how-uk-pays-success.
9. Ibid.

References

Gaylor, Erika, Traci Kutaka, Kate Ferguson, Cyndi Williamson, Xin Wei, and Donna Spiker. 2016. "Evaluation of Kindergarten Readiness in Five Child-Parent Centers: Report for 2014–15." Menlo Park, CA: SRI International.

Gibbs, Graham R. 2007. *Analyzing Qualitative Data*. London: SAGE Publications.

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/.

Harrell, Adele. 1996. "Evaluation Strategies for Human Services: A Guide for Policymakers and Providers." Washington, DC: Urban Institute. https://www.bja.gov/evaluation/guide/documents/evaluation_strategies.html.

Milner, Justin, and Kelly Walsh. 2016. "Measuring Success in Pay for Success: Randomized Controlled Trials as the Starting Point." Washington, DC: Urban Institute. http://www.urban.org/research/publication/measuring-success-pay-success-randomized-controlled-trials-starting-point.

Nonprofit Finance Fund. 2016. "Pay for Success: The First Generation." New York: Nonprofit Finance Fund. http://www.payforsuccess.org/sites/default/files/Pay%20for%20Success_The%20First%20Generation.pdf.

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

Winkler, Mary K., Brett Theodos, and Michel Grosz. 2009. *Evaluation Matters: Lessons from Youth-Serving Organizations*. Washington, DC: Urban Institute. http://www.urban.org/research/publication/evaluation-matters.

About the Authors

Kelly A. Walsh is a senior research associate in the Justice Policy Center at the Urban Institute, where her work focuses on social science research in forensic science and innovative social financing arrangements such as social impact bonds or pay for success financing. She is most interested in questions centered on the efficacy of forensic processes, the causes of wrongful convictions, and the mechanisms of private investment for the sake of public good. She earned a BS in chemistry from the University of Scranton and her PhD in criminal justice, with a specialization in forensic science, from the Graduate Center, City University of New York.

Rebecca TeKolste is a former project associate with the Urban Institute's Pay for Success Initiative. Before her time at Urban, she served as a Peace Corps Volunteer in rural Guatemala, working on a maternal and child health project. She is currently pursuing an MA in global affairs at Yale University's Jackson Institute.

Ben Holston is a former research assistant in the Policy Advisory Group at the Urban Institute. He is a graduate of Stanford University and currently resides in the San Francisco Bay Area.

John K. Roman is a former senior fellow in the Justice Policy Center at the Urban Institute. His research focuses on evaluations of innovative crime control policies and justice programs. He has directed several studies funded by the National Institute of Justice, including two randomized trials of the costs and benefits of using of DNA in motor vehicle thefts and burglary investigations, a study investigating why forensic evidence is rarely used by law enforcement to identify unknown offenders, and a study on wrongful conviction. Roman is a lecturer at the University of Pennsylvania and an affiliated professor at Georgetown University.

Acknowledgments

This brief was funded by the Laura and John Arnold Foundation. We are grateful to them and to all 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 Akiva Liberman and Janine Zweig for reviewing an earlier version of this brief.

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.